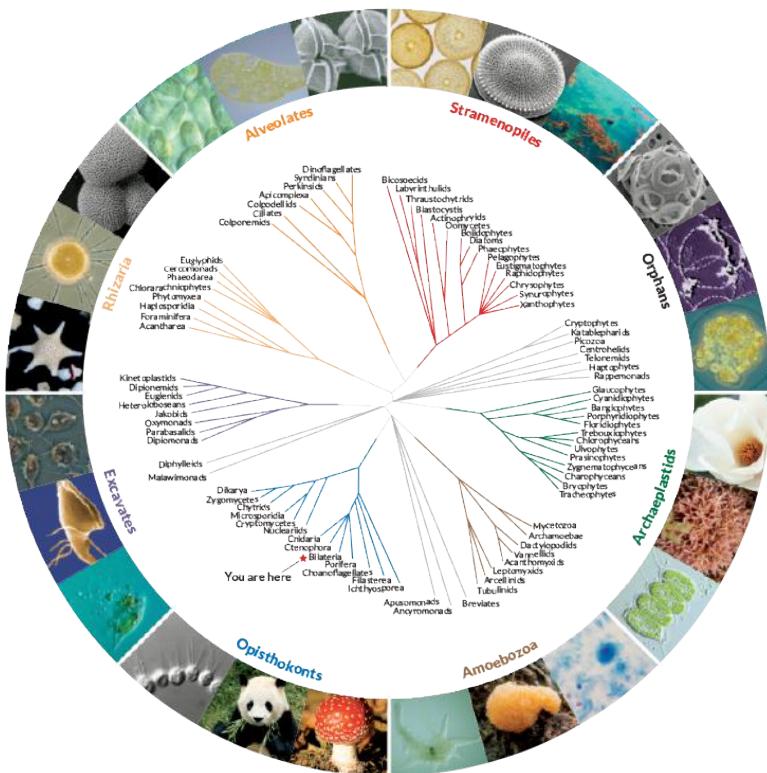


Instituto Português do Mar e da Atmosfera

Guia técnico de curso de formação

Biologia Molecular e Bioinformática



Bárbara Frazão e Teixeira João Paulo Machado

Créditos da imagem: A. Worden et al/Science 2015



Introdução à Biologia Molecular e à Bioinformática

1^a edição: 1 a 3 de Junho de 2016

2^a edição: 12 a 14 de Outubro de 2016

Curso BIOMAR PT
2016

Formadores

Bárbara Frazão

João Paulo Machado

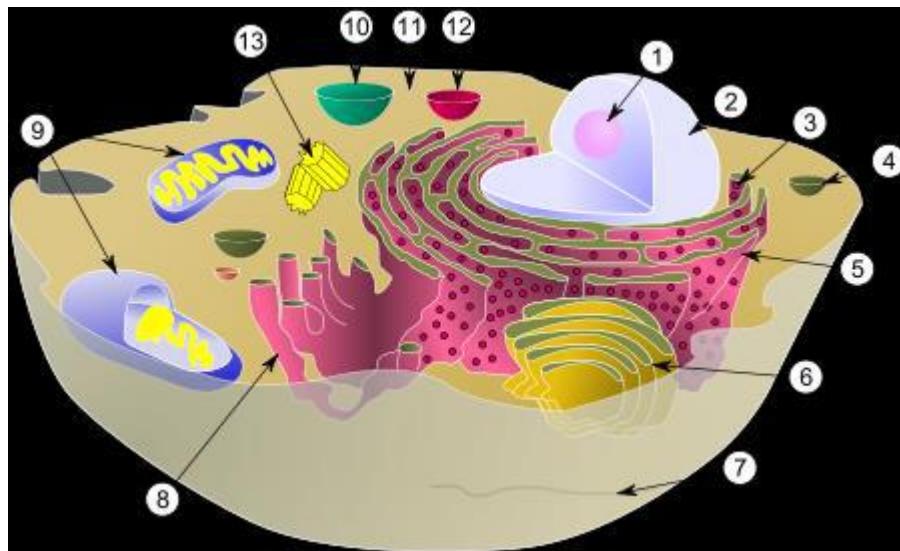


Estrutura do curso

- **1^a aula (T) 4h - BM**
Conceitos básicos de Biologia molecular, preparação das amostras, Primer design, PCR, sequenciação
- **2^a aula (P) 3h –BM**
Extração DNA genómico, Electroforese em gel de agarose e Quantificação de DNA genómico
- **3^a aula (T) 4h- BI**
 - Bases de dados, alinhamentos, modelos evolutivos, filogenias
- **4^a aula (P) 3h- BM**
 - PCR e corrida em gel
- **5^a aula (TP) 3h – BI**
 - Desenho de primers, BLAST, alinhamentos, determinação do modelo evolutivo
- **6^a aula (TP) 3h- BI**
 - Construção de Filogenias

Conceitos básicos de Biologia molecular

- A **Biologia Molecular** é o estudo da Biologia a nível molecular, com ênfase na estrutura e função do material genético e seus produtos de expressão, as proteínas.



- Animal cell
 - cytoplasm
 - (1) nucleolus
 - (2) nucleus
 - (3) ribosome
 - (4) vesicle
 - (5) rough endoplasmic reticulum
 - (6) Golgi apparatus
 - (7) cytoskeleton
 - (8) smooth endoplasmic reticulum
 - (9) mitochondria
 - (10) vacuole
 - (11) cytosol
 - (12) lysosome
 - (13) centriole.

Conceitos básicos de Biologia molecular

DNA

Descoberto por James Watson (norte-americano) e Francis Crick (britânico) em 1953

Prémio Nobel de Fisiologia ou Medicina em 1962

Ácido desoxirribonucleico

Composto orgânico cujas moléculas contêm as instruções genéticas que coordenam o desenvolvimento e funcionamento de todos os seres vivos e alguns vírus, e que transmitem as características hereditárias de cada ser vivo.

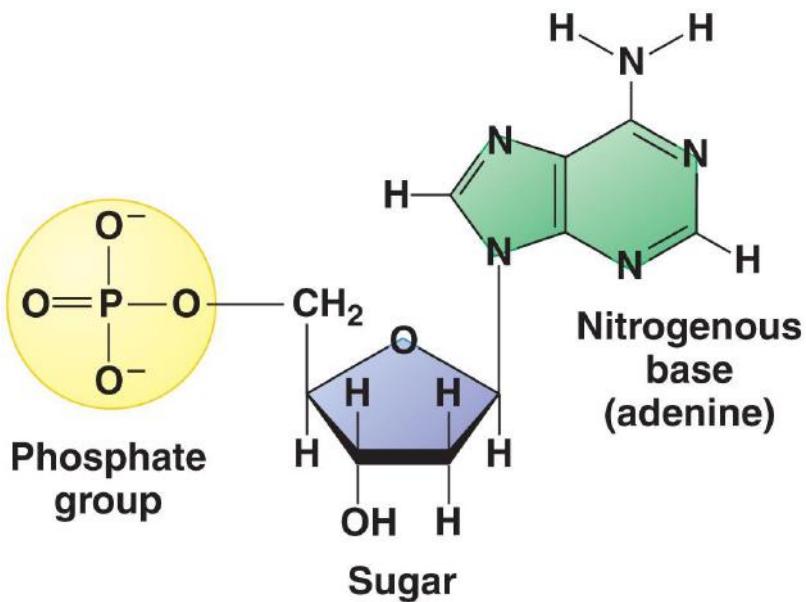
O seu principal papel é armazenar as informações necessárias para a construção das proteínas de RNAs.



Conceitos básicos de Biologia molecular

Ácidos Nucleicos

Polímero de unidades simples
(monómeros)-
os nucleótidos

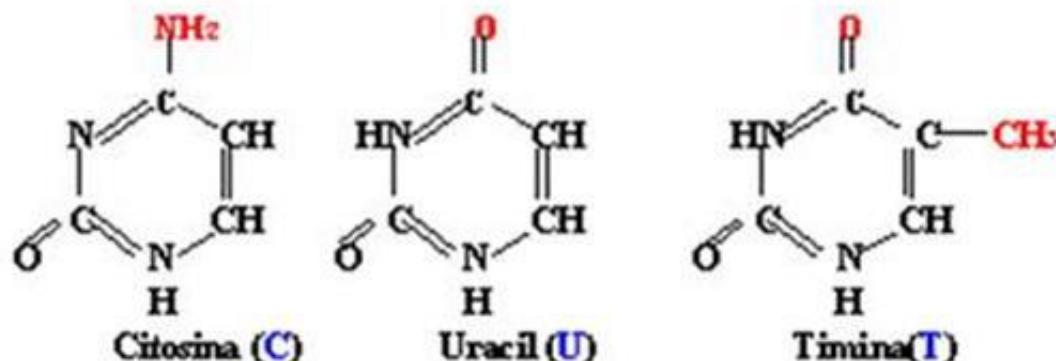
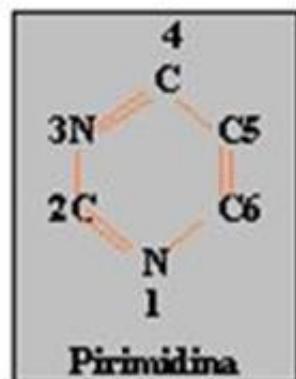
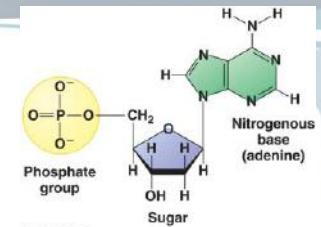
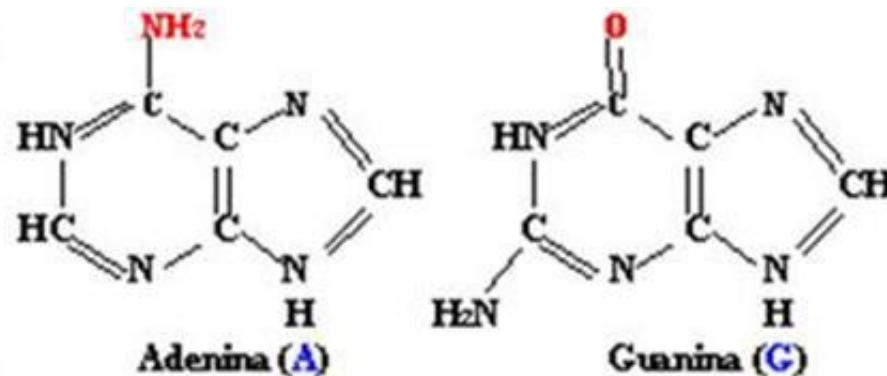
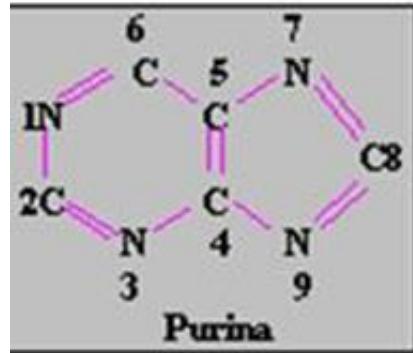


Copyright © 2009 Pearson Education, Inc.

Conceitos básicos de Biologia molecular

Ácidos Nucleicos

5 tipos de bases azotadas



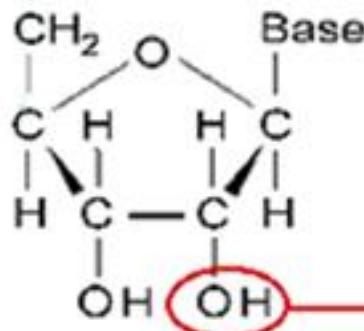
Conceitos básicos de Biologia molecular

Ácidos Nucleicos

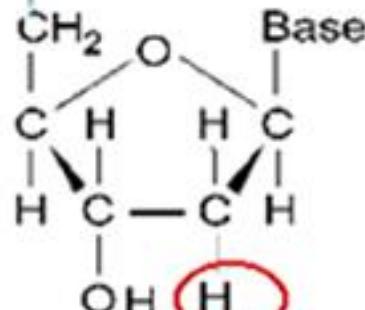
2 tipos de Pentoses

Ribose

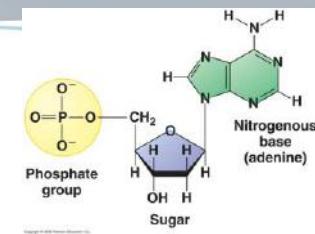
Fosfato



Fosfato

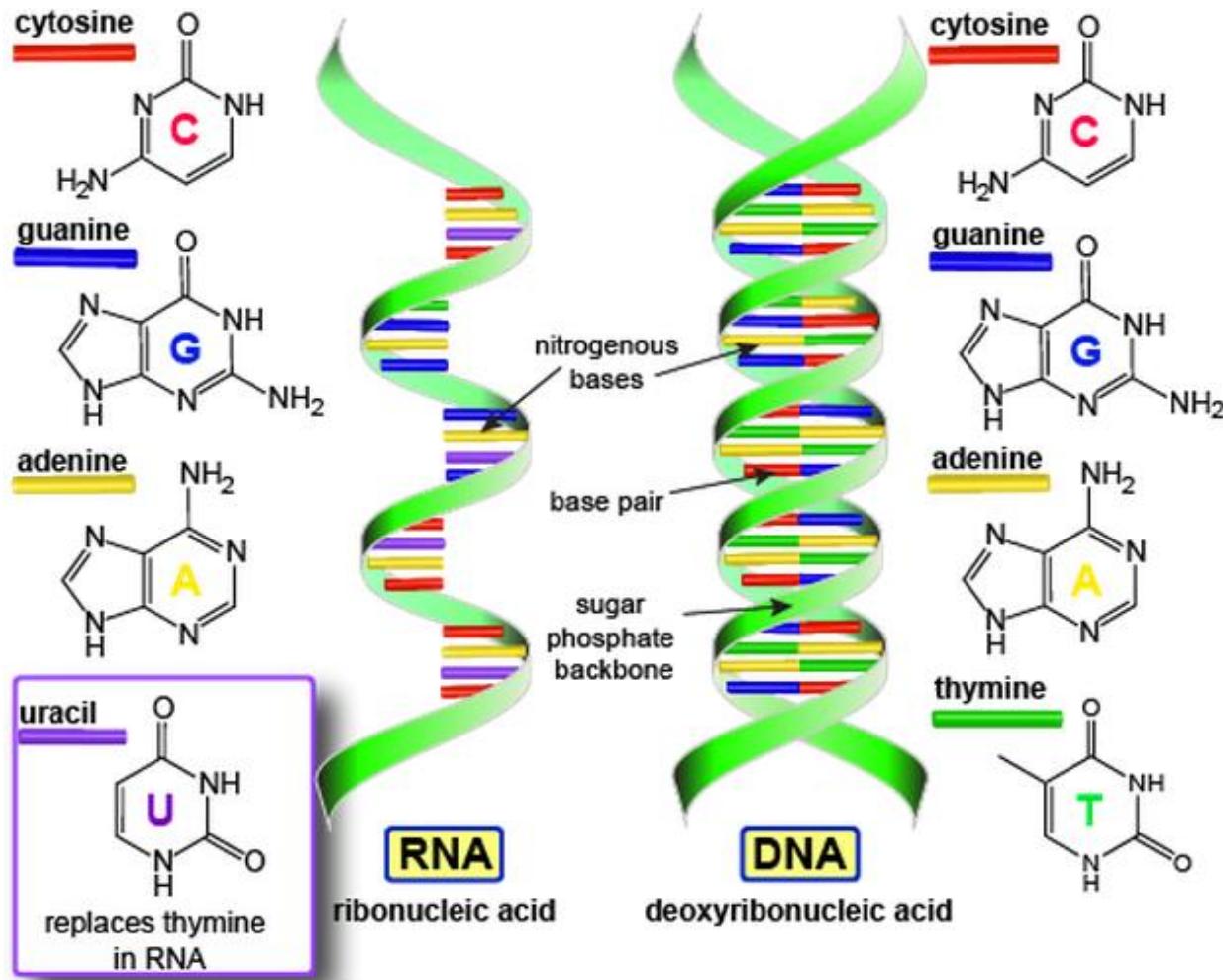


Desoxirribose



Conceitos básicos de Biologia molecular

Ácidos Nucleicos



Conceitos básicos de Biologia molecular

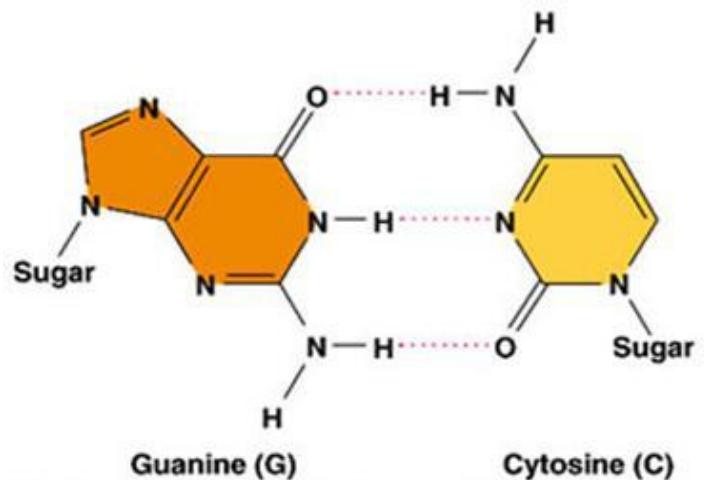
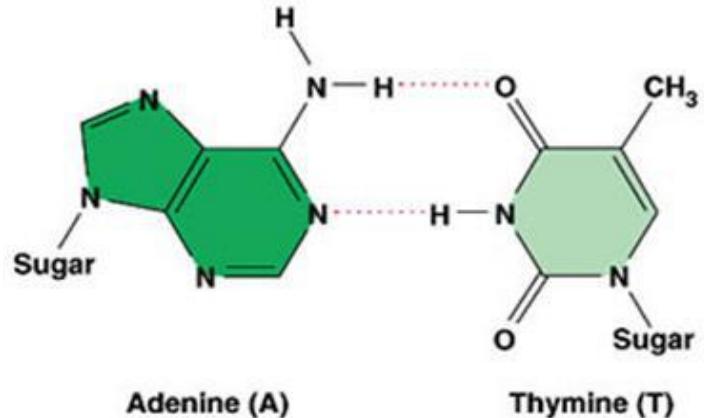
Ácidos Nucleicos

	DNA	RNA
Pentose	Desoxirribose	Ribose
Bases púricas	Adenina e Guanina	Adenina e Guanina
Estrutura	Citosina e Timina	Citosina e Uracila
Bases primídicas	Duas cadeias helicoidais	Uma cadeia
Enzima hidrolítica	Desoxirribonuclease (DNAase)	Ribonuclease (RNAase)
Origem	Replicação	Transcrição
Enzima sintética	DNA - polimerase	RNA - polimerase
Função	Informação genética	Síntese de proteínas

Conceitos básicos de Biologia molecular

DNA

- Adenina complementar Timina
 - Ligação por 2 pontes de hidrogénio
- Citosina complementar com Guanina
 - Ligação por 3 pontes de hidrogénio

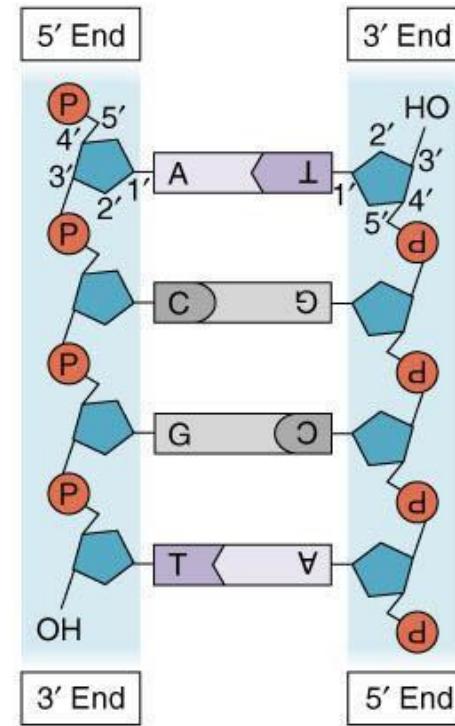
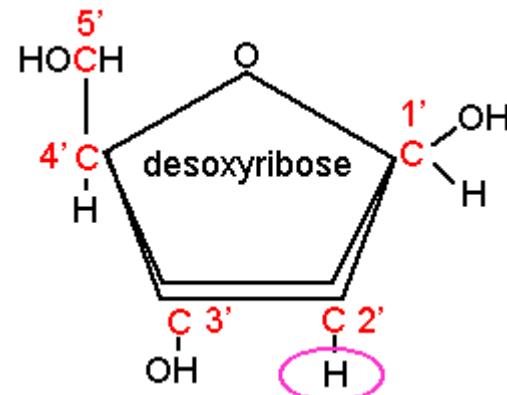
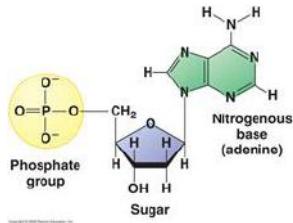


Copyright © Pearson Education, Inc., publishing as Benjamin Cummings.

Conceitos básicos de Biologia molecular

DNA

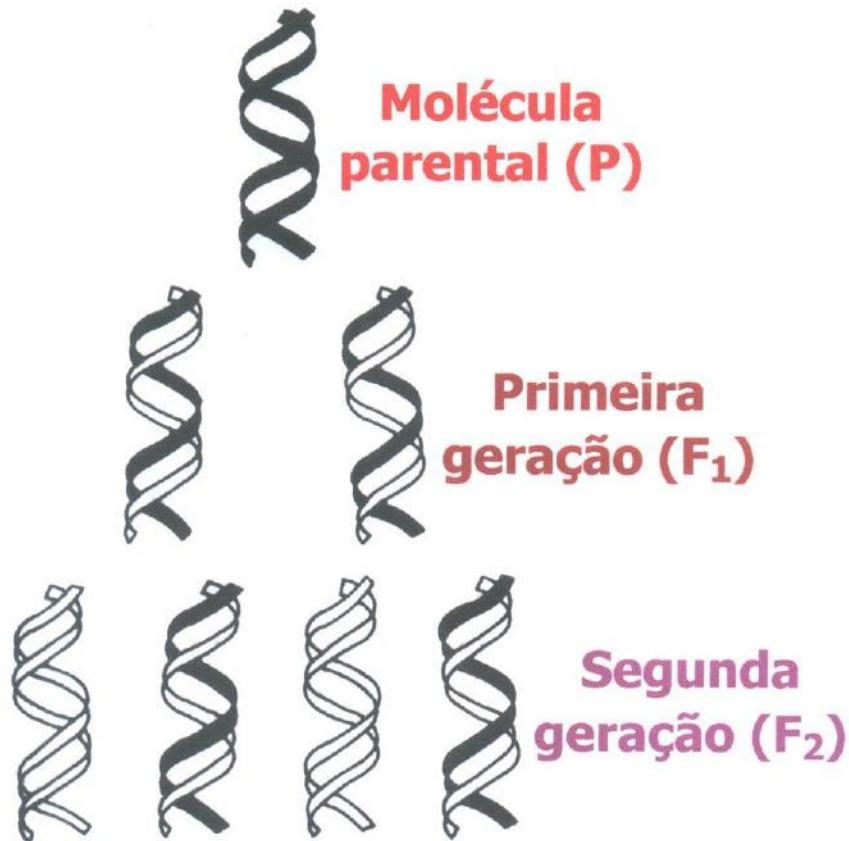
- 2 cadeias helicoidais complementares e antiparalelas



- Extremo 3' – tem um grupo hidroxilo no 3º carbono
- Extremo 5' - tem um grupo fosfato no 5º carbono
- Numeração dos carbonos- de 1' a 5' por convenção no sentido dos ponteiros do relógio, a partir do oxigénio
- Importância: síntese dos ácidos nucleicos dá-se no sentido 3'

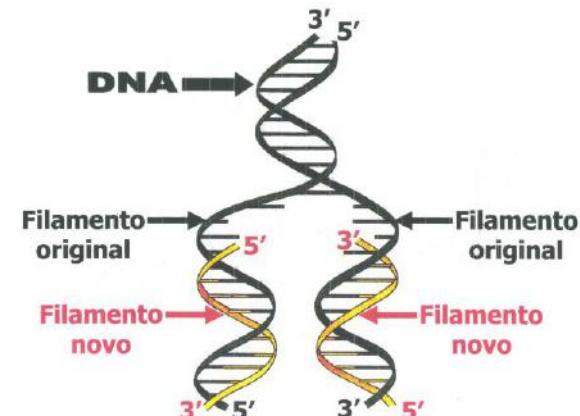
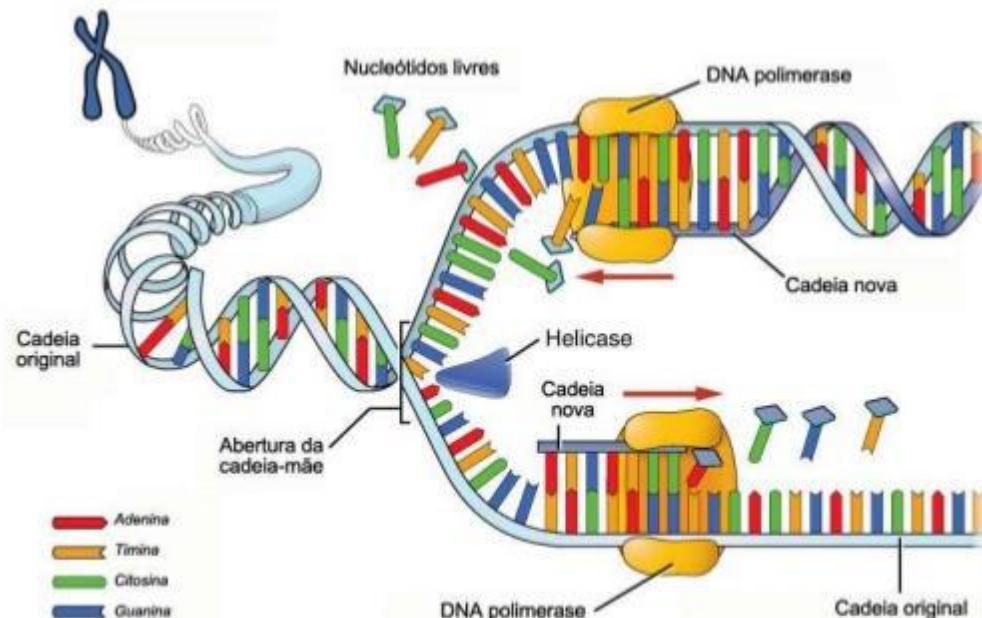
Conceitos básicos de Biologia molecular

Replicação semi-conservativa do DNA



Conceitos básicos de Biologia molecular

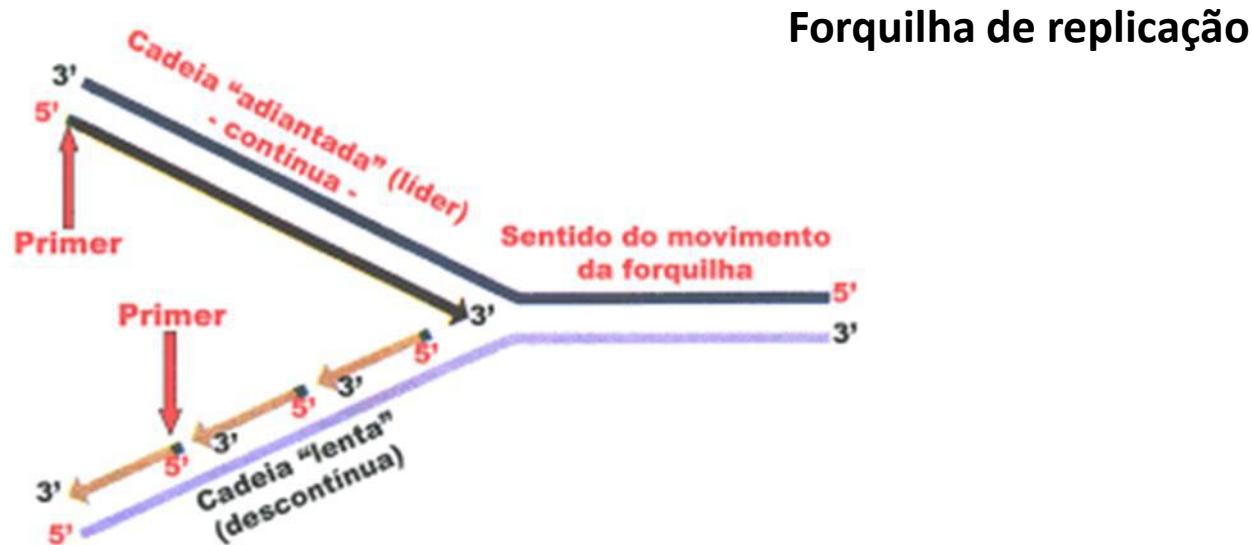
Replicação semiconservativa do DNA



- 1. Helicases-** separam as 2 cadeias da molécula
- 2. DNA Polimerase III-** catalizam a adição de um nucleótido ao radical hidroxilo na extremidade 3' da cadeia que se está formando. Desta forma as cadeias só podem crescer no sentido 5' → 3'

Conceitos básicos de Biologia molecular

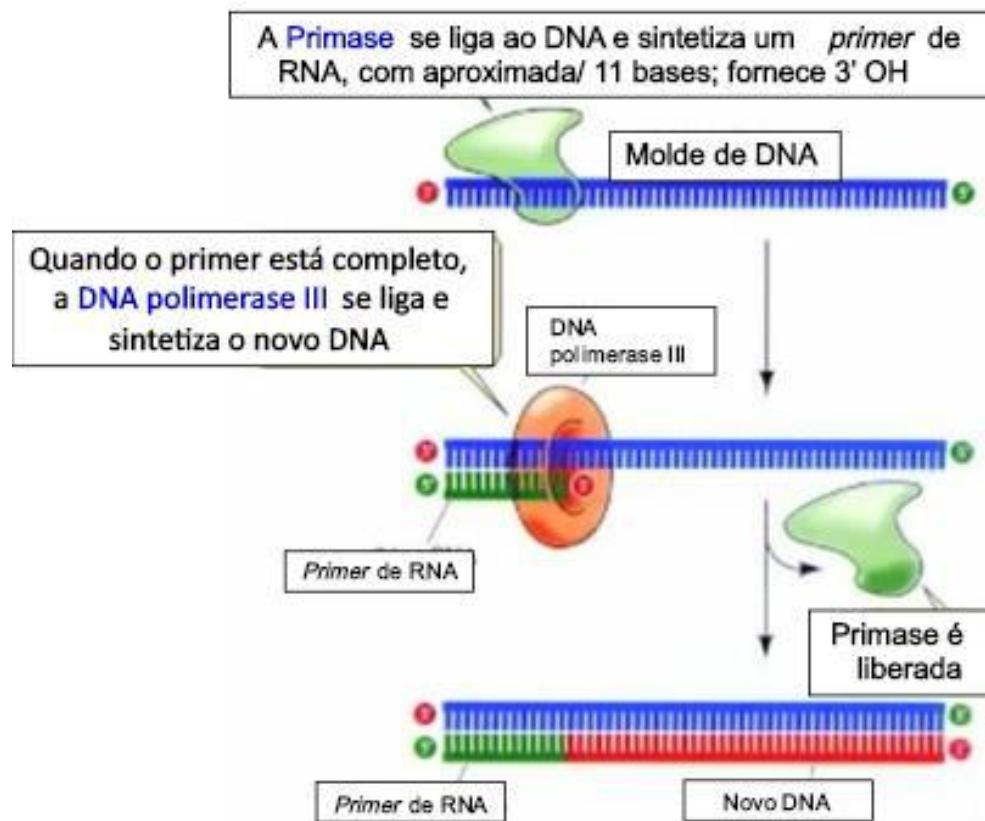
Replicação semiconservativa do DNA



- **Cadeia contínua** - sintetizada continuamente a partir de um iniciador na cadeia molde 3' → 5'
- **Cadeia descontínua** – sintetizada descontinuamente a partir de múltiplos iniciadores

Conceitos básicos de Biologia molecular

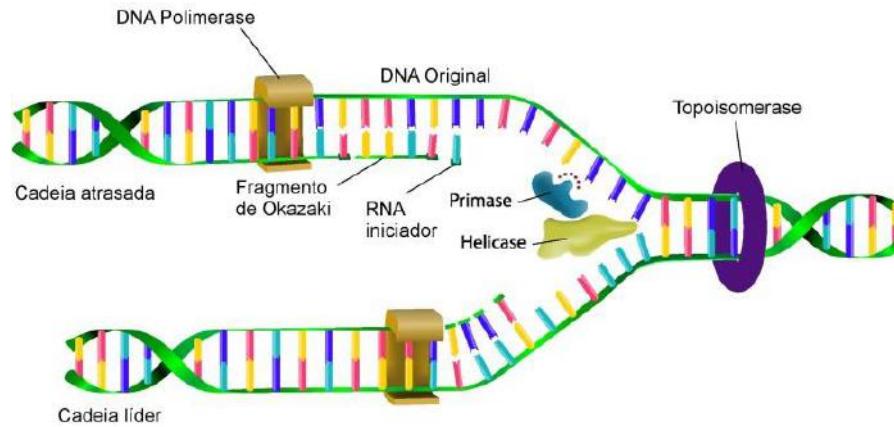
Replicação semiconservativa do DNA



3. Primase- Sintetiza pequenas moléculas de RNA utilizadas como iniciadores durante o processo de replicação

Conceitos básicos de Biologia molecular

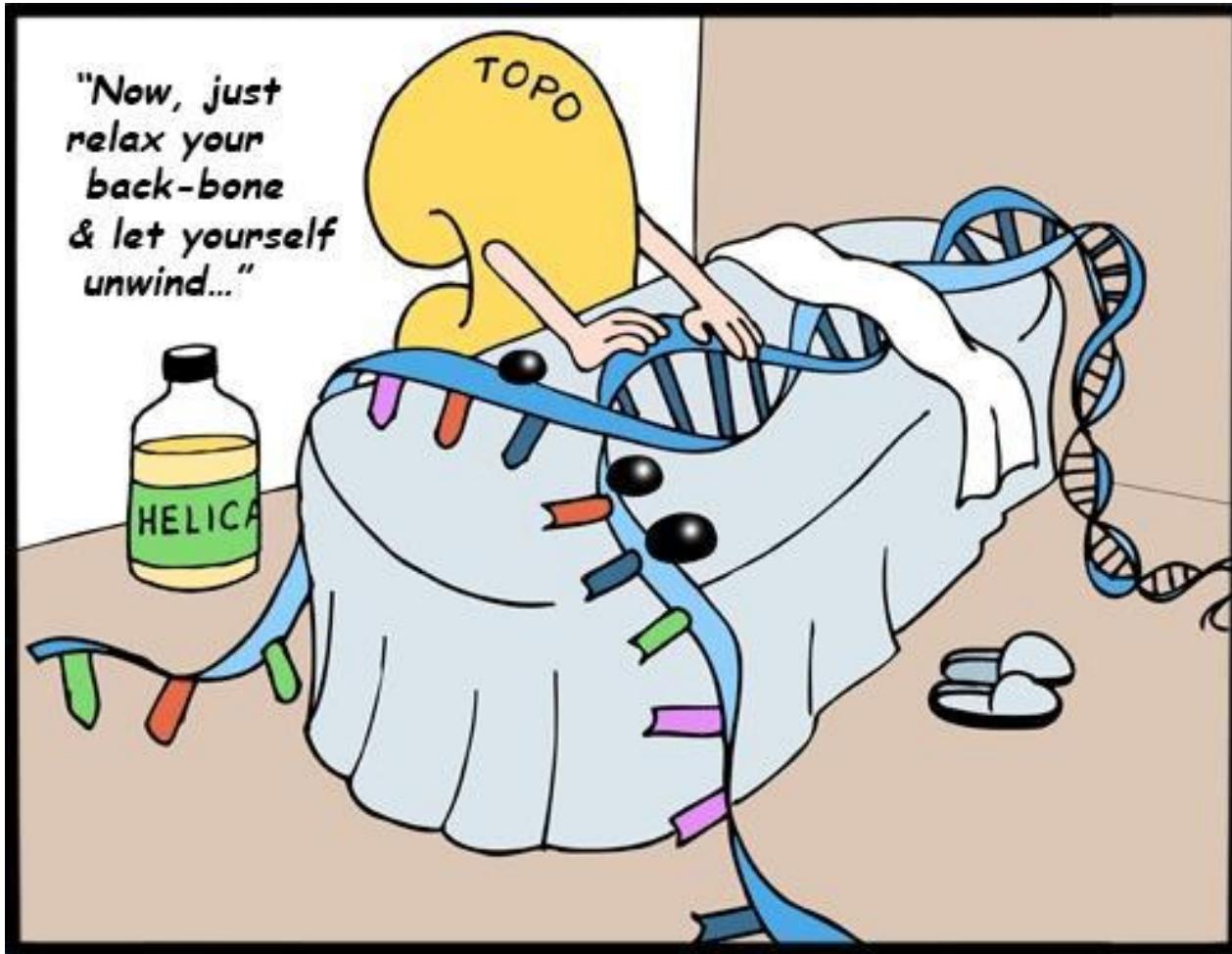
Replicação semiconservativa do DNA



Topoisomerase
alivia a torção na
parte da cadeia
que não está a ser
replicada

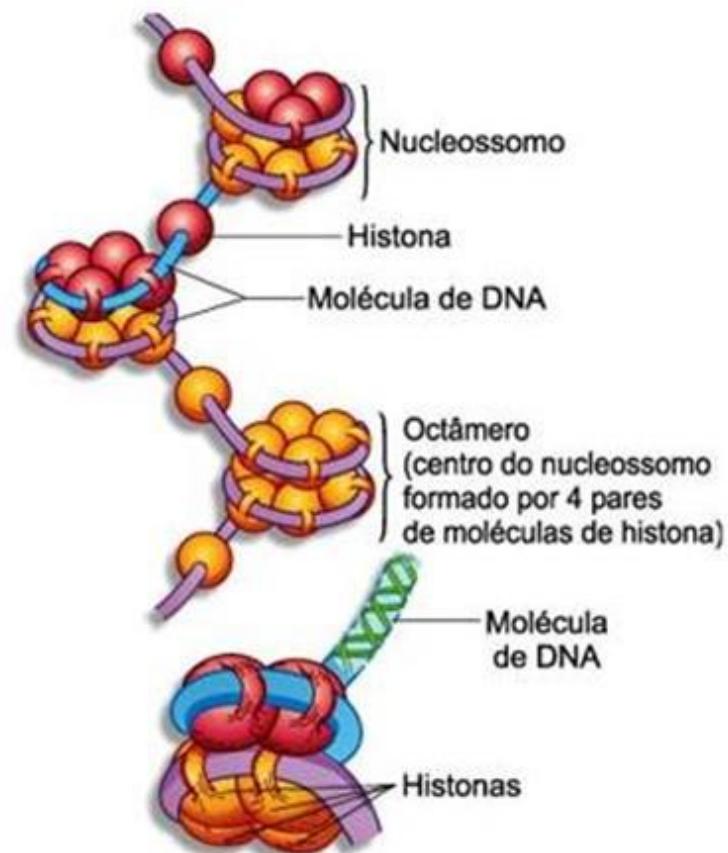
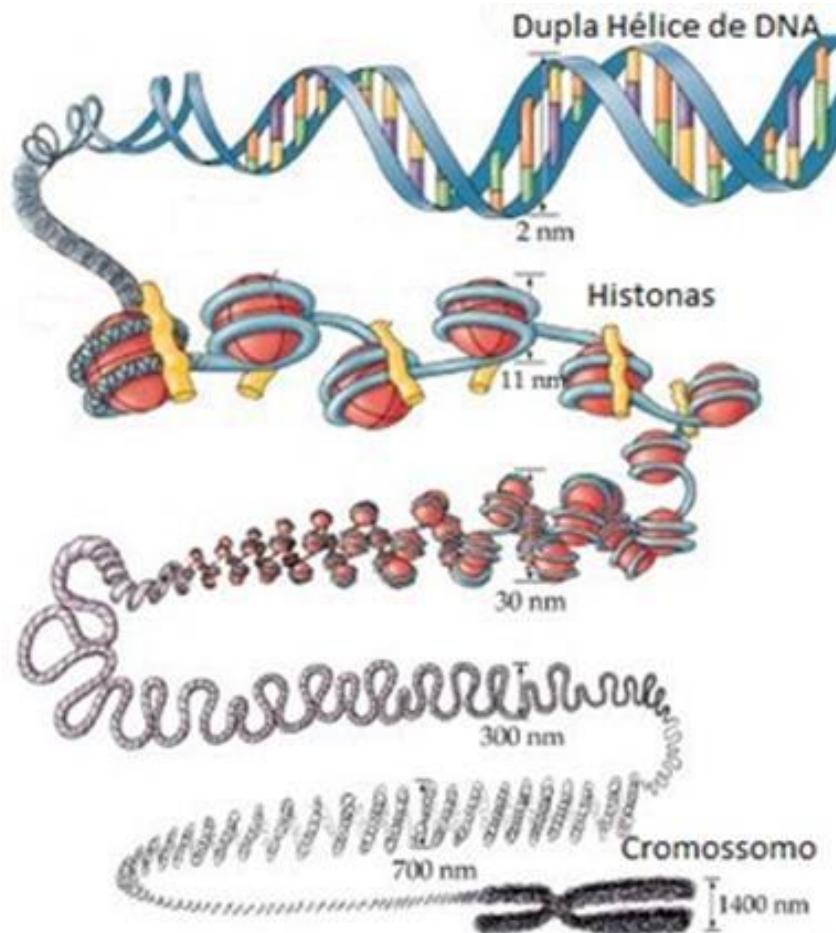
- Cada iniciador é alongado pela DNA polimerase resultando na formação de **Fragmentos de Okazaki**
- **DNA polimerase I** remove o primer do RNA do fragmento adjacente e preenche os gaps entre os fragmentos
- **DNA ligase** – liga os fragmentos de Okasaki

Conceitos básicos de Biologia molecular



Conceitos básicos de Biologia molecular

Organização do DNA

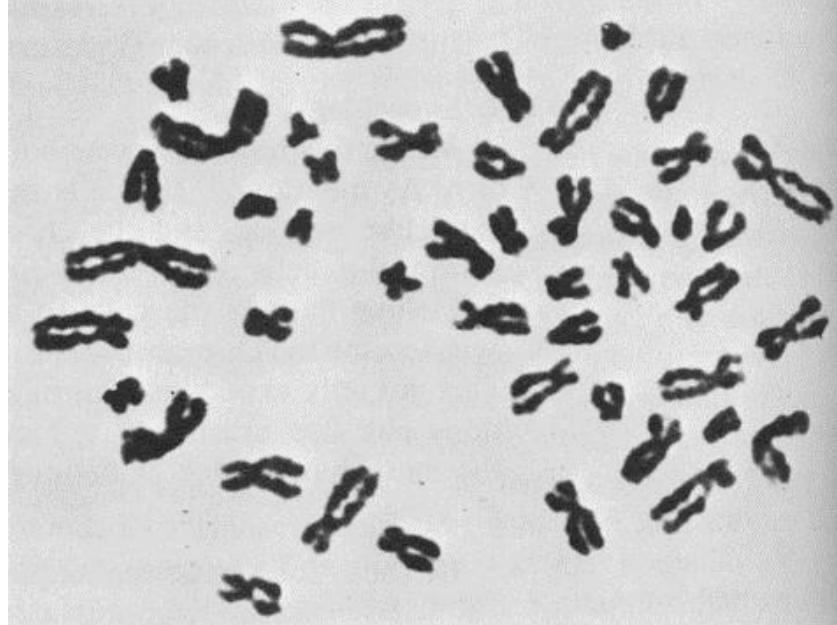
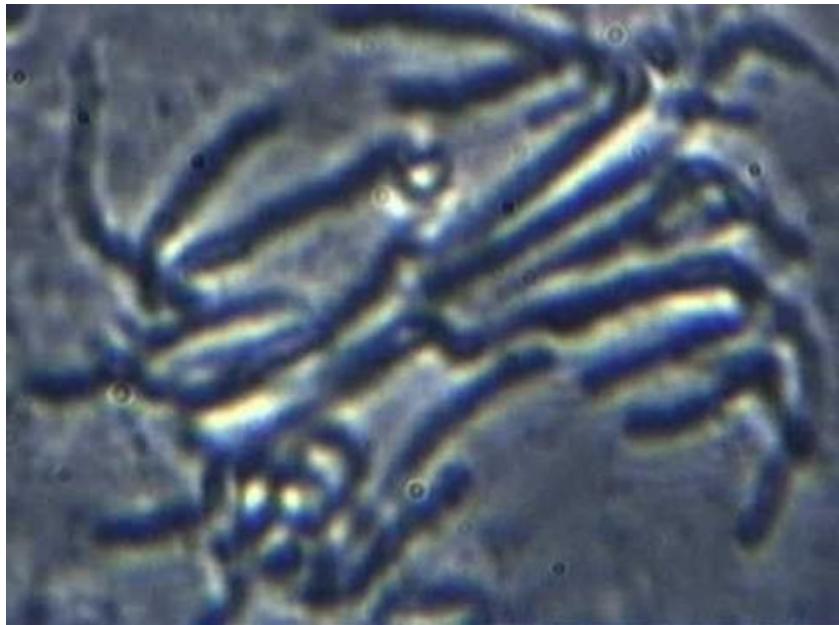


A organização molecular da cromatina.

Adaptado
© 1998 Sinauer Associates, Inc.

Conceitos básicos de Biologia molecular

Cromossomas

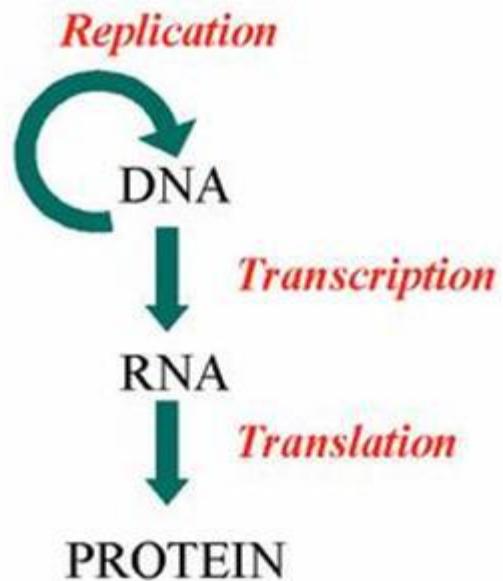


Cariótipo

Conjunto de cromossomas de uma célula

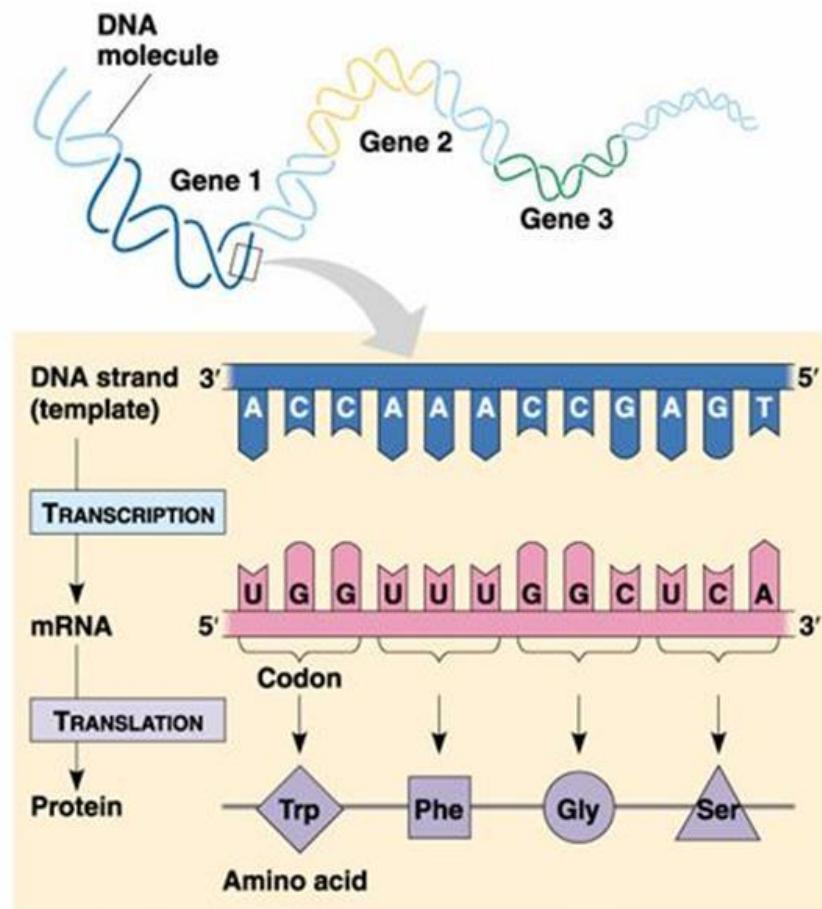
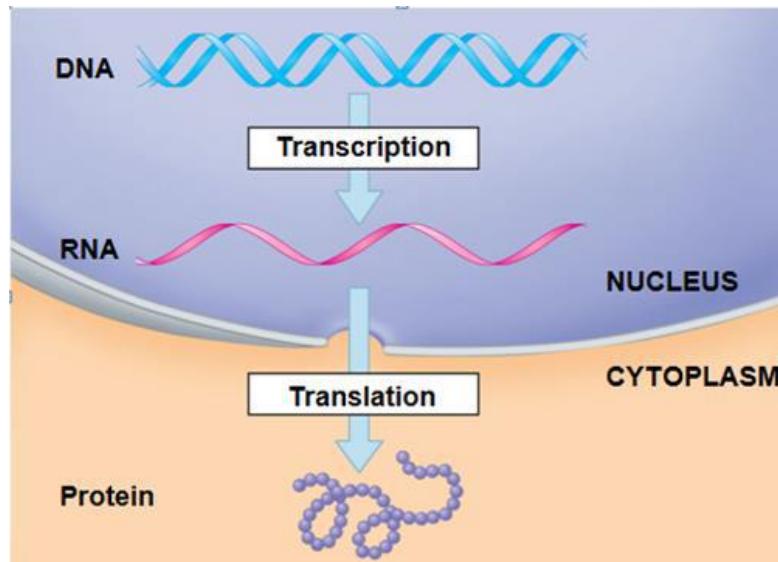
Conceitos básicos de Biologia molecular

Transcrição e Tradução



Conceitos básicos de Biologia molecular

Transcrição e Tradução



©1999 Addison Wesley Longman, Inc.

Conceitos básicos de Biologia molecular

Transcrição e Tradução

A **transcrição** é a passagem da informação genética, do núcleo para o citoplasma, sob a forma de **RNA mensageiro (mRNA)**, cuja sequencia é complementar ao DNA de que foi transcrito



Messenger RNA
Carries instructions for polypeptide synthesis from nucleus to ribosomes in the cytoplasm.



Ribosome

Ribosomal RNA
Forms an important part of both subunits of the ribosome.



Amino acid

Transfer RNA
Carries amino acids to the ribosome and matches them to the coded mRNA message.

Tipos de RNA

- Mensagem
- Ribossómico
- de Transferência

Conceitos básicos de Biologia molecular

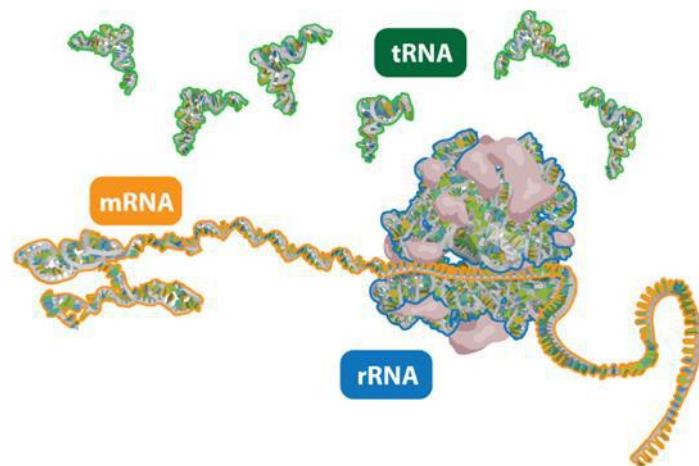
Código genético

Gene- região de DNA que codifica para determinada proteína ou RNA funcional

Genoma- todo o material genético de um organismo e inclui quer os genes quer as sequencias não-codificantes

Sequências não codificantes, e.g.

- RNA não codificantes
 - transferência RNA
 - ribosómico RNA
 - regulatório RNA
- Regulação da atividade das regiões codificantes



Conceitos básicos de Biologia molecular

Código genético

A sequência de bases ao longo da molécula de ADN constitui a informação genética. A leitura destas sequências é feita por intermédio do código genético, que especifica a sequência linear dos aminoácidos das proteínas.

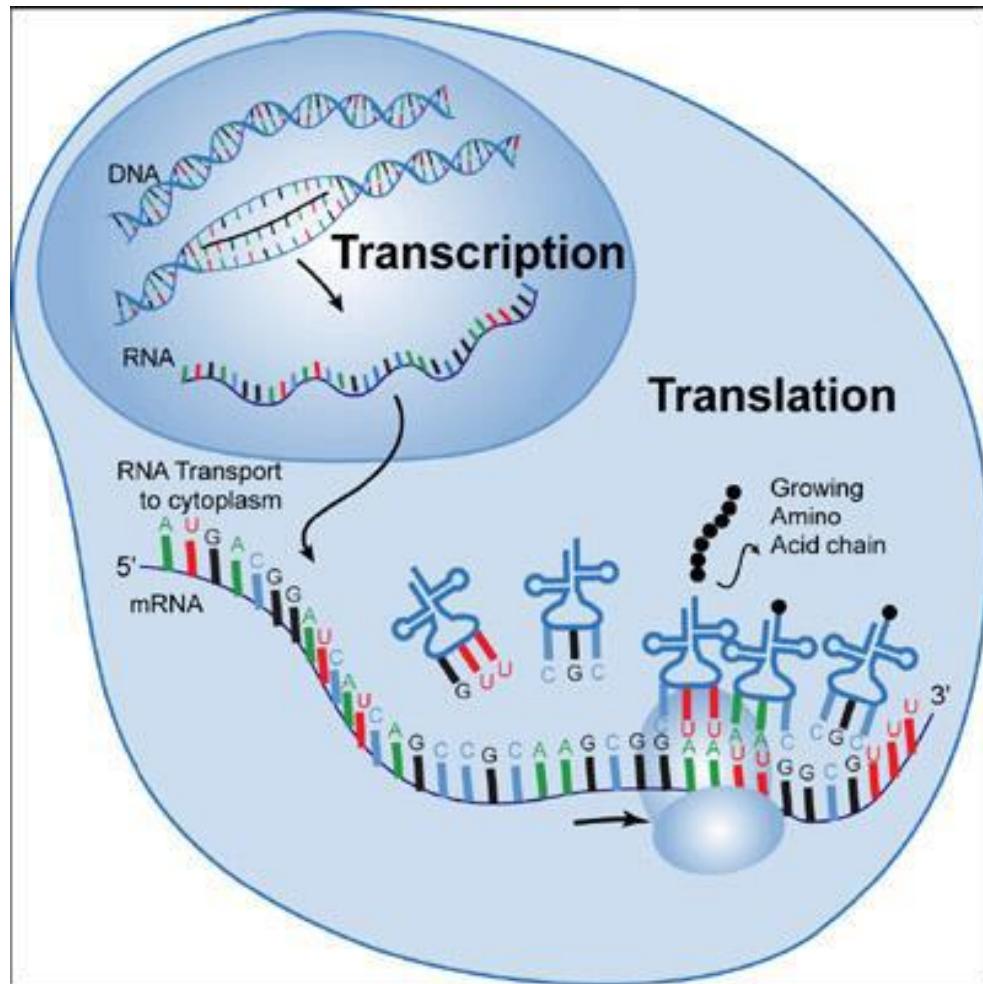
		Second letter				Third letter
		U	C	A	G	
First letter	U	UUU } Phe UUC UUA } Leu UUG }	UCU } Ser UCC UCA UCG }	UAU } Tyr UAC UAA Stop UAG Stop }	UGU } Cys UGC UGA Stop UGG Trp }	U C A G
	C	CUU } Leu CUC CUA CUG }	CCU } Pro CCC CCA CCG }	CAU } His CAC CAA } Gin CAG }	CGU } Arg CGC CGA CGG }	U C A G
A	AUU } Ile AUC AUA } Met AUG }	ACU } Thr ACC ACA ACG }	AAU } Asn AAC AAA } Lys AAG }	AGU } Ser AGC AGA } Arg AGG }	U C A G	
G	GUU } Val GUC GUA GUG }	GCU } Ala GCC GCA GCG }	GAU } Asp GAC GAA } Glu GAG }	GGU } Gly GGC GGA GGG }	U C A G	

Codão- 3 nucleótidos seguidos do mRNA

Anticodão- as 3 bases do tRNA, que são complementares ao codão do mRNA

Conceitos básicos de Biologia molecular

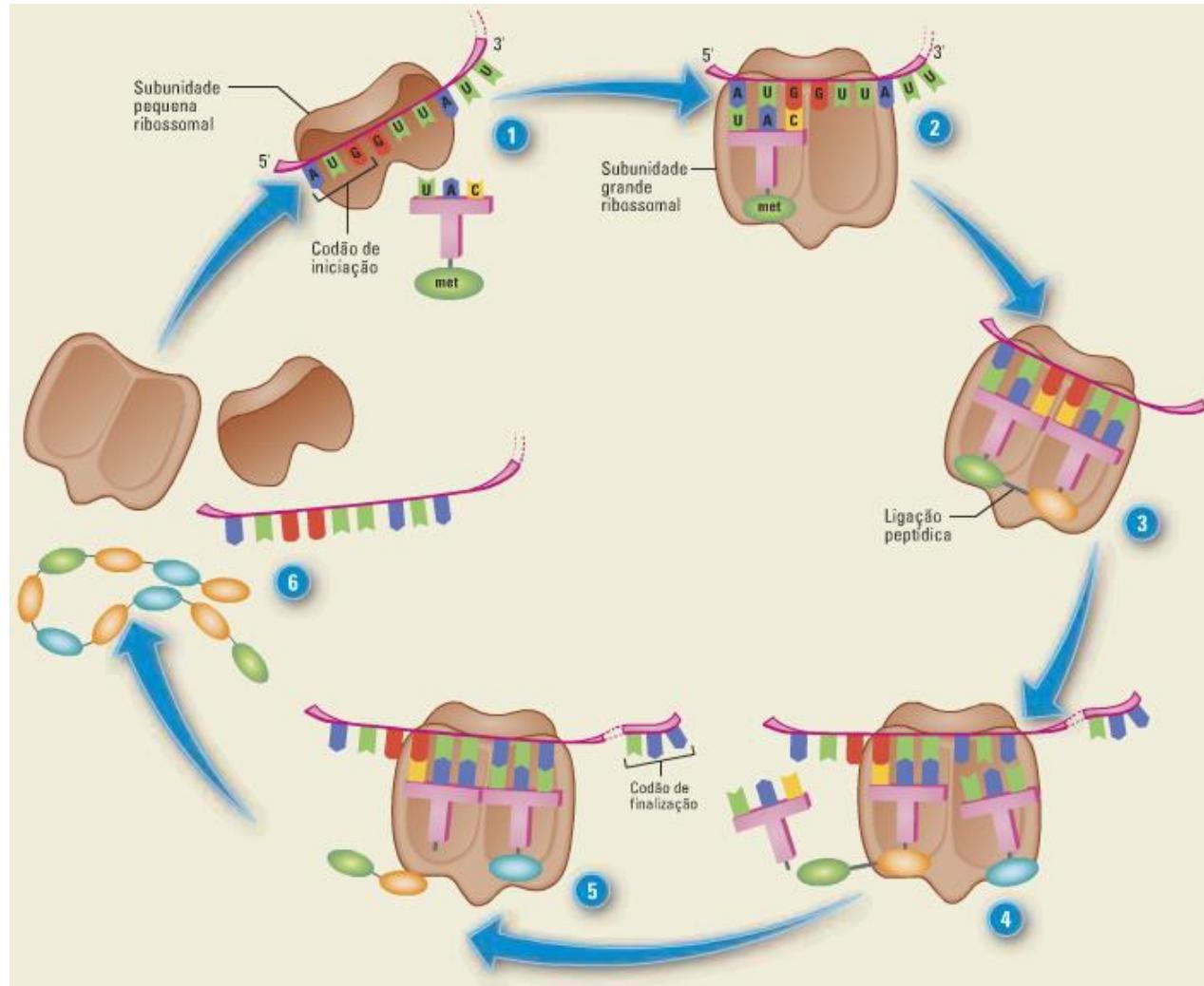
Tradução



Processo pelo qual o RNA maduro, serve de “template” para a síntese de uma nova proteína

Conceitos básicos de Biologia molecular

Tradução



Extração de DNA genómico

Método tradicional Fenólico

Ref: <http://dx.doi.org/10.2216/i0031-8884-42-3-261.1>

1. 50ml culture
 2. Centrifugue and discard supernatant
 3. Pellet resuspended in 750ul filtered NSW
 4. 3500g, 10min, 4°C
 5. Pellet stored in liquid nitrogen
-
1. Nitrogen pump – 10 cycles, 2400psi (to lyse)
 2. 500ul lysis buffer at 65°C, 30min
 - 0.02M EDTA
 - 2% CTAB
 - 0.1M Tris
 - 1.4M NaCl
 - 9.2% 2-b-mercaptoetanol

Extração de DNA genómico

Método tradicional Fenólico

Lysis Buffer- permite quebrar as células para libertar o DNA, sem a sua degradação

Podem conter:

1. Sais-manter uma força iônica (concentração salina)
2. Detergentes- separar as proteínas da membrana
 - Triton X-100
 - CHAPS
 - SDS
3. RNase A – degrada RNA
4. Proteinase K- desnatura proteínas, mas geralmente tem um tempo definido de atuação

Extração de DNA genómico

Método tradicional Fenólico

3. 100ul phenol-chloroform-isoamyl alcohol (25:24:1)
4. 14000g 8min, 4°C
5. Repeat step 3 com o sobrenadante
6. Chloroform-isoamyl alcohol (24:1)
7. 14000g 8min, 4°C
8. 5ul glycogen ao sobrenadante
9. 15ul 3M NaOAc and isopropanol
10. Incubate -80°C, 30min
11. 16000g, 15min
12. 1000ul, 70%ethanol
13. 16000g, 15min
14. Dried in a speed vacum
15. Ressuspend 50ul H₂O

Extração de DNA genómico

Método tradicional : Fenólico

Phenol-chloroform-isoamyl alcohol

DNA extraction proporção 25:24:1

pH DNA extraction 6.7-8

(RNA – pH fenol 4.8)

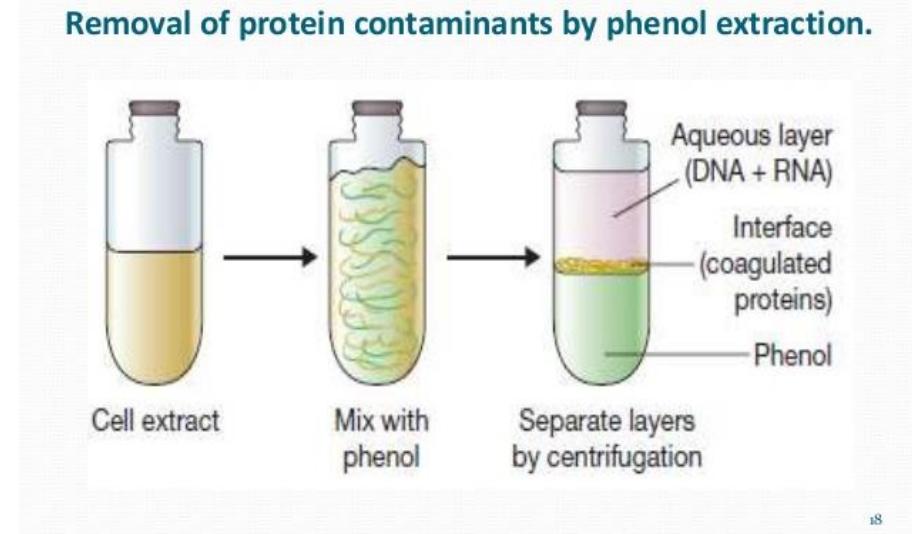
Clorofórmio- desnatura proteínas

Isoamyl alcohol- previne formação de espuma

Sais de Guanidina – reduzem os efeitos das nucleases

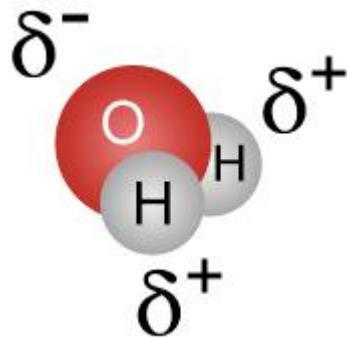
Glicogénio é insolúvel no etanol e forma um precipitado que agarra o ácido nucleico

Fenol ácido retém o RNA na fase aquosa, mas move o DNA para a fase fenólica

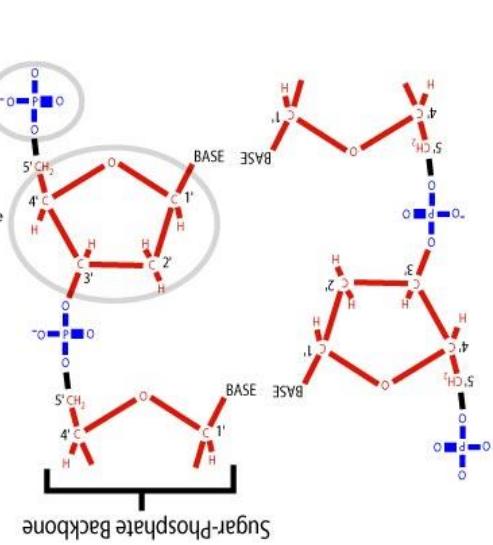
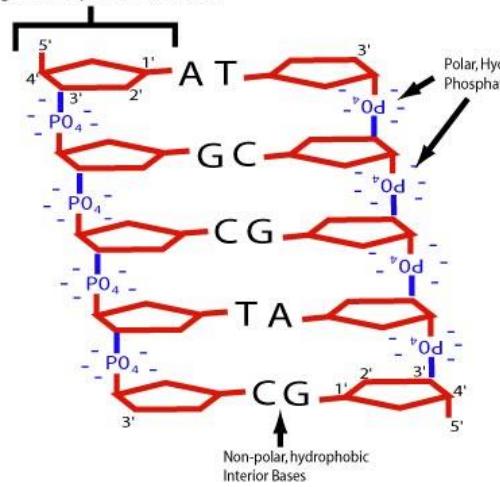


Extração de DNA genómico

Método tradicional : Fenólico



Sugar-Phosphate Backbone



DNA rodeado por moléculas de água

Extração de DNA genómico

Método tradicional : Fenólico

Precipitação de DNA

1. Sal para neutralizer a carga do ácido nucleico. O DNA fica menos hidrofílico e precipita na solução (acetato de sódio, cloreto de sódio, cloreto de litio, acetate de amónio, etc)
2. Etanol 100% para precipitar o DNA (altera a estrutura do DNA de maneira em que as suas moléculas se agregam e precipitam)
3. Frio para arrefecer a amostra. Temperaturas baixas promovem a floculação dos ácidos nucleicos, para que formem um complex maior que facilite a formação de pellet na centrifugação (Over-night -4°C ou -80°C 1 hora)
4. Uma concentração elevada de ácido nucleico, suficiente para forçar recuperação do DNA (se não for elevada pode adicionar-se um “carrier” como o glicogénio para aumentar a recuperação)
5. Centrifugação da amostra até formar um pellet

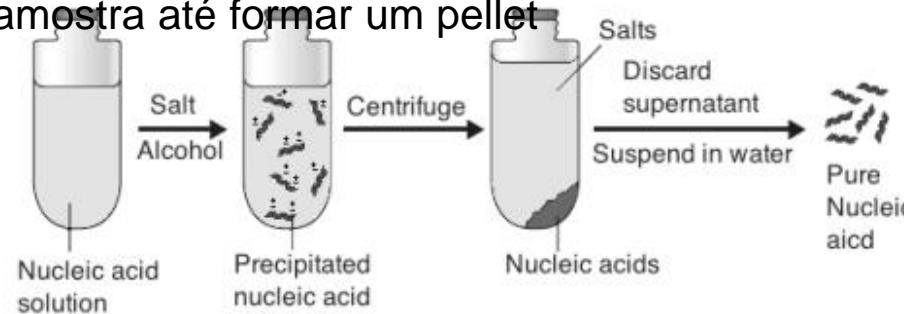


Figure 1. Schematic overview of an ethanol precipitation of nucleic acids.

Extração de DNA genómico

Método tradicional : Fenólico

Etanol	Isopropanol
Se couberem 2 volumes de etanol	Se couber apenas 1 volume de álcool (tubo pequeno)
Precipita fragmentos menores e maiores	Precipita apenas fragmentos maiores
Precipita a -20ºC	Precipita a RT
Precipita pouca concentração de amostra	Só precipita se em quantidade considerável de amostra
Precipitação demorada	Precipitação rápida

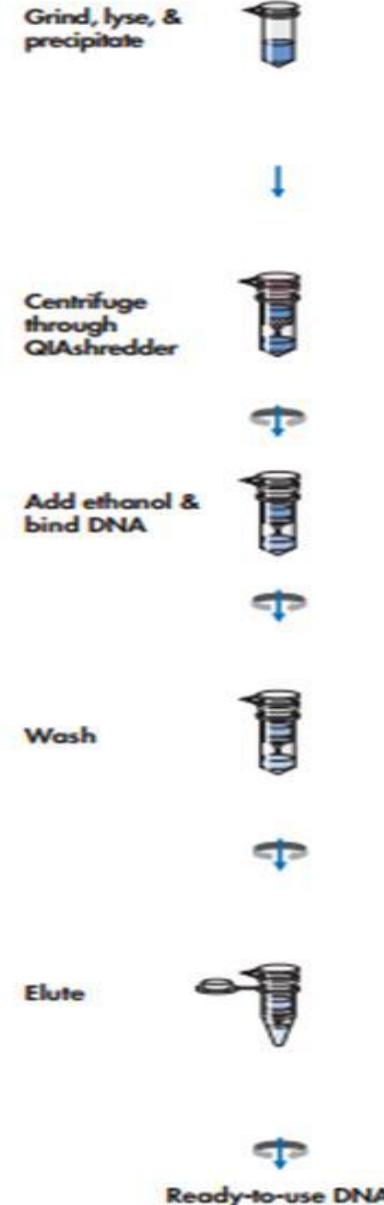
Lavagem do pellet com etanol a 70% ethanol. Percentagem na qual os sais são removidos e o DNA não é dissolvido

Secagem do pellet:

- Deve ser eficaz, para o etanol não estar presente na amostra
- Não deve ser demasiado, pois dificulta a dissolução na solução de eluição

Extração de DNA genómico

Kit da Qiagen- DNeasy plant



Extração de DNA genómico

Kit da Qiagen- DNeasy plant

1. Add 400 µl Buffer AP1 and 4 µl RNase A and vortex
2. Homogeneização da amostra no miniLys- 15s, 5000rpm

Extração de DNA genómico

Kit da Qiagen-DNeasy plant

- Mais volume amostra no mesmo volume de lysis buffer/Coluna, não aumenta eficiência de extração

8. Incubate the mixture for 10 min at 65°C. Mix 2 or 3 times during incubation by inverting tube.

This step lyses the cells.

9. Add 130 µl Buffer P3 to the lysate, mix, and incubate for 5 min on ice.

This step precipitates detergent, proteins, and polysaccharides.

10. Recommended: Centrifuge the lysate for 5 min at 20,000 x g (14,000 rpm).

Some plant materials can generate very viscous lysates and large amounts of precipitates during this step. This can result in shearing of the DNA in the next step (see "Lysate filtration with QIAshredder", page 19). In this case, optimal results are obtained if the majority of these precipitates are removed by centrifugation for 5 min at 20,000 x g (14,000 rpm). After centrifugation, apply supernatant to QIAshredder Mini spin column and continue with step 11.

11. Pipet the lysate into the QIAshredder Mini spin column (lilac) placed in a 2 ml collection tube, and centrifuge for 2 min at 20,000 x g (14,000 rpm).

It may be necessary to cut the end off the pipet tip to apply the lysate to the QIAshredder Mini spin column. The QIAshredder Mini spin column removes most precipitates and cell debris, but a small amount will pass through and form a pellet in the collection tube. Be careful not to disturb this pellet in step 12.

12. Transfer the flow-through fraction from step 11 into a new tube (not supplied) without disturbing the cell-debris pellet.

Typically 450 µl of lysate is recovered. For some plant species less lysate is recovered. In this case, determine the volume for the next step.

13. Add 1.5 volumes of Buffer AW1 to the cleared lysate, and mix by pipetting.

For example, to 450 µl lysate, add 675 µl Buffer AW1. Reduce the amount of Buffer AW1 accordingly if the volume of lysate is smaller. A precipitate may form after the addition of Buffer AW1, but this will not affect the DNeasy procedure.

Note: Ensure that ethanol has been added to Buffer AW1. See "Things to do before starting", page 22.

Note: It is important to pipet Buffer AW1 directly onto the cleared lysate and to mix immediately.

Extração de DNA genómico

Kit da Qiagen-DNeasy plant

14. Pipet 650 µl of the mixture from step 13, including any precipitate that may have formed, into the DNeasy Mini spin column placed in a 2 ml collection tube (supplied). Centrifuge for 1 min at $\geq 6000 \times g$ (corresponds to ≥ 8000 rpm for most microcentrifuges), and discard the flow-through.* Reuse the collection tube in step 15.
15. Repeat step 14 with remaining sample. Discard flow-through* and collection tube.

* Flow-through fractions contain Buffer AW1, and are therefore not compatible with bleach. See page 6 for safety information.

24

DNeasy Plant Handbook 06/2015

16. Place the DNeasy Mini spin column into a new 2 ml collection tube (supplied), add 500 µl Buffer AW2, and centrifuge for 1 min at $\geq 6000 \times g$ (≥ 8000 rpm). Discard the flow-through and reuse the collection tube in step 17.

Note: Ensure that ethanol is added to Buffer AW2. See "Things to do before starting", page 22.

17. Add 500 µl Buffer AW2 to the DNeasy Mini spin column, and centrifuge for 2 min at $20,000 \times g$ (14,000 rpm) to dry the membrane.

It is important to dry the membrane of the DNeasy Mini spin column since residual ethanol may interfere with subsequent reactions. This centrifugation step ensures that no residual ethanol will be carried over during elution. Discard flow-through and collection tube.

After washing with Buffer AW2, the DNeasy Mini spin column membrane is usually only slightly colored. In the rare case that the membrane remains significantly colored after washing with Buffer AW2, refer to "Darkly colored membrane or green/yellow eluate after washing with Buffer AW2" in the Troubleshooting Guide on page 43.

Note: Following the centrifugation, remove the DNeasy Mini spin column from the collection tube carefully so the column does not come into contact with the flow-through, as this will result in carryover of ethanol.

Extração de DNA genómico

Kit da Qiagen-DNeasy plant

- Eluição em 30ul, optimiza quantidade de DNA
- Secagem total de etanol, Incubação RT (room temperature) e Elution Buffer 50-60°C também optimiza quantidade de DNA

18. Transfer the DNeasy Mini spin column to a 1.5 ml or 2 ml microcentrifuge tube (not supplied), and pipet 100 µl Buffer AE directly onto the DNeasy membrane. Incubate for 5 min at room temperature (15–25°C), and then centrifuge for 1 min at $\geq 6000 \times g$ (≥ 8000 rpm) to elute.

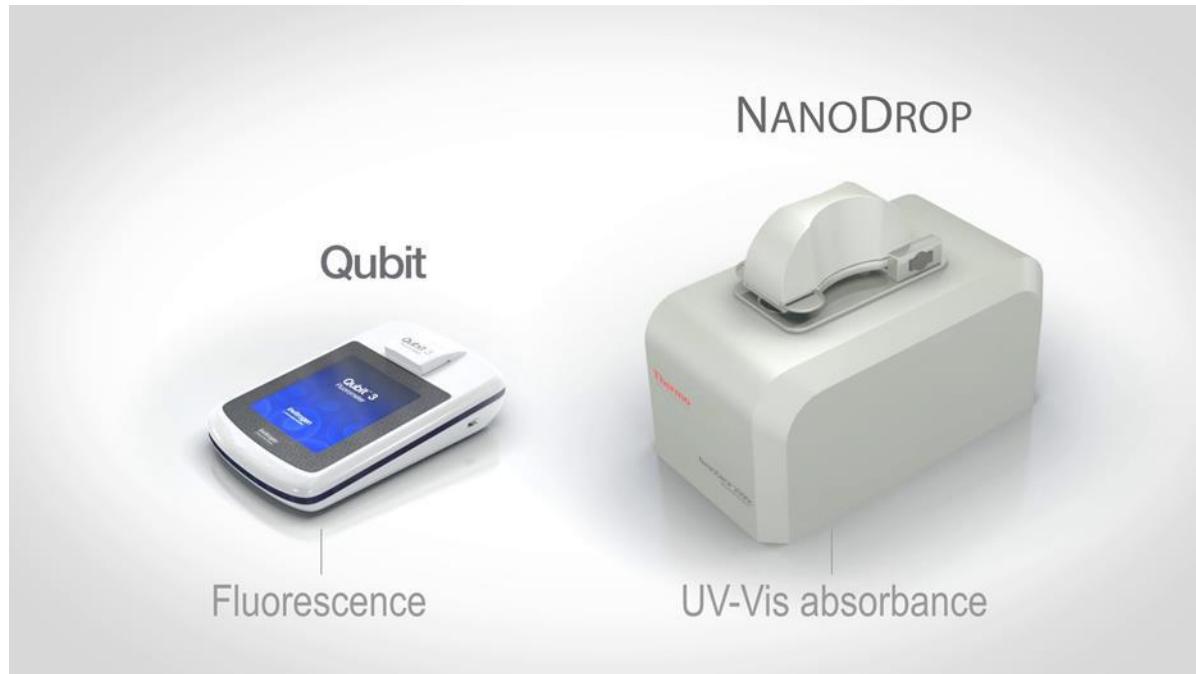
Elution with 50 µl (instead of 100 µl) increases the final DNA concentration in the eluate significantly, but also reduces overall DNA yield. If larger amounts of DNA (>20 µg) are loaded, eluting with 200 µl (instead of 100 µl) increases yield. See "Elution", page 19.

19. Repeat step 18 once.

A new microcentrifuge tube can be used for the second elution step to prevent dilution of the first eluate. Alternatively, the microcentrifuge tube can be reused for the second elution step to combine the eluates. See "Elution", page 19.

Note: More than 200 µl should not be eluted into a 1.5 ml microcentrifuge tube because the DNeasy Mini spin column will come into contact with the eluate.

Quantificação do DNA



Quantificação do DNA

Table 3. Quantification method comparison.

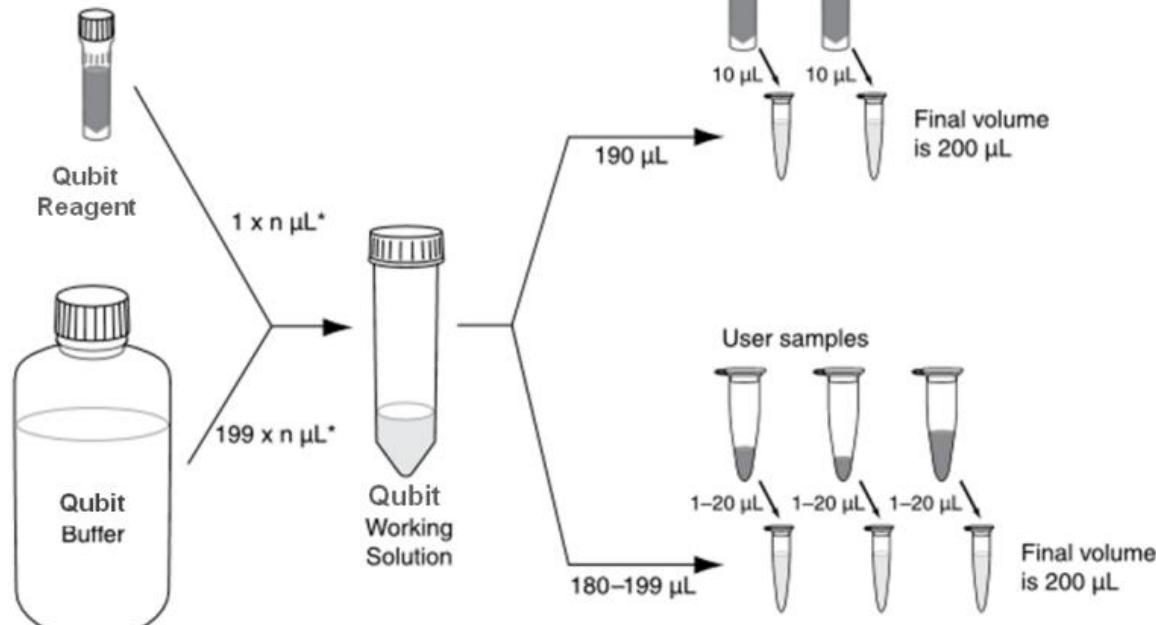
	Qubit® Fluorometer	UV-absorbance microvolume spectrophotometer
Quantification method	Fluorescence-based dyes that bind specifically to DNA, RNA, or protein	UV absorbance measurements (measures absorbance at 260 nm and 260 nm/280 nm ratio)
Accuracy and precision at low concentrations (Figure 4)	Accurately quantifies DNA in samples with concentrations as low as 10 pg/µL	Not recommended for concentrations under 2 ng/µL; variation for sample concentrations <10 ng/µL is often high
Sensitivity and range (Figure 5)	The effective range covers a sample concentration range of 10 pg/µL to 1 µg/µL DNA	Covers a sample concentration range of 2 ng/µL to 15 µg/µL; uses 0.5–2 µL of sample
Can indicate contamination	No	Gives peaks revealing the presence of contaminants

Quantificação do DNA

Product	Cat. No.	No. of assays	Quantitation range	Applications
Qubit dsDNA BR Assay Kit	Q32850	100	2–1,000 ng	Quantitation of genomic and miniprep DNA samples
	Q32853	500		
Qubit dsDNA HS Assay Kit	Q32851	100	0.2–100 ng	Quantitation of PCR products, viral DNA, or samples for NGS
	Q32854	500		

Quantificação do DNA

Ensure all reagents are at room temperature



* where n = number of standards plus number of samples

Vortex all assay tubes for 2–3 seconds

Incubate at room temperature for 2 minutes

Read tubes in Qubit™ fluorometer



Quantificação do DNA



- Pipetar 2ul de amostra, sem criar bolhas e sem tocar no equipamento
- Ler a absorvância do “Branco”- Solução na qual o DNA está eluído
- Ler a absorvância da nossa amostra
- Abs 260/280 deve estar entre 1.8 e 2. Valores >2 indicam contaminação com RNA. Valores inferiores a 1.8 indicam contaminação com Fenol ou proteínas, por exemplo.

Quantificação do DNA

Quantificação de DNA genómico em gel de agarose

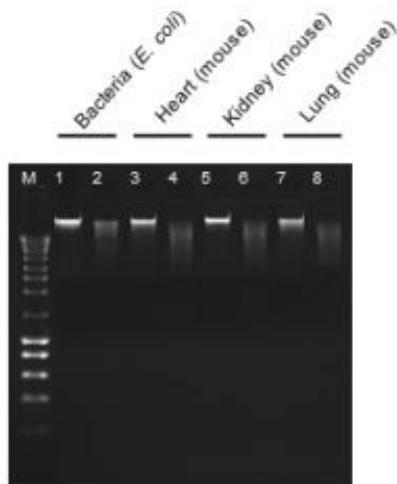
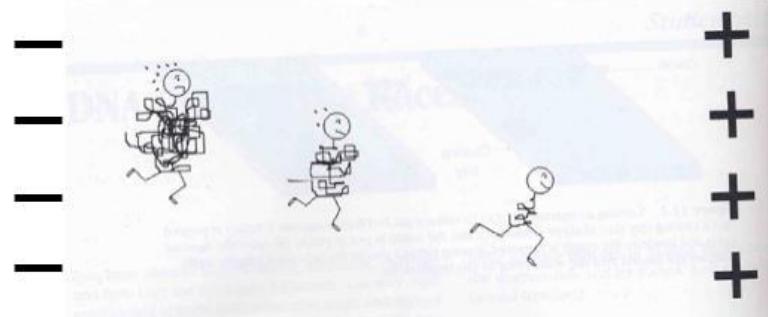
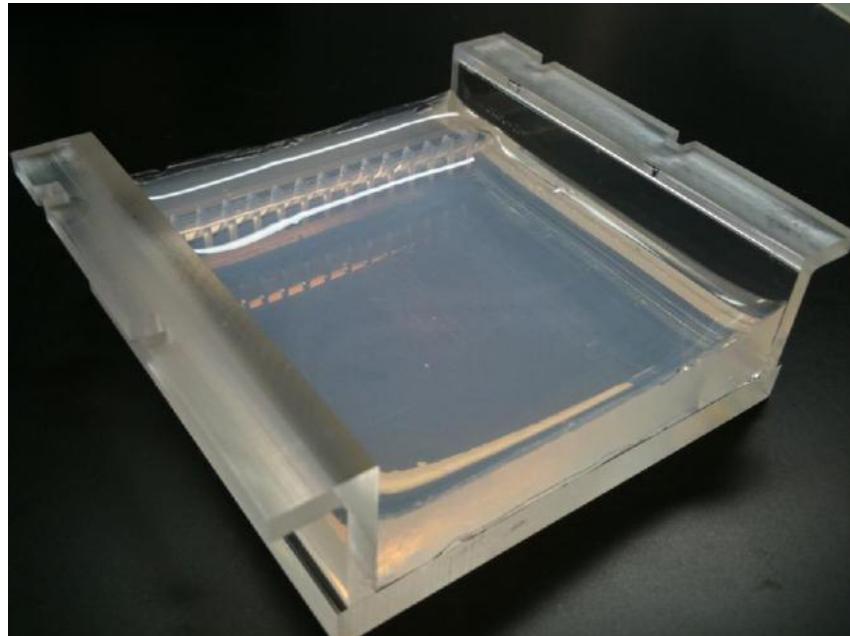


Fig. 1. Genomic DNA from various sources isolated with NZY Tissue gDNA Isolation kit. 0.25 µg of each isolated DNA was analysed on a 0.8% agarose gel. Lanes 2, 4, 6 and 8: gDNA digested with *EcoR* I. Lanes 1, 3, 5 and 7: undigested gDNA. M: NZYDNA Ladder III (MB0440); 1, 2: gDNA from bacteria (*E. coli*); 3, 4: gDNA from mouse heart; 5, 6: gDNA from mouse kidney; 7, 8: gDNA from mouse lung.

Electroforese em gel de agarose

A eletroforese em gel é uma técnica de separação de moléculas que envolve a migração de partículas num determinado gel durante a aplicação de uma diferença de potencial. As moléculas são separadas de acordo com o seu tamanho, pois as de menor massa irão migrar mais rapidamente que as de maior massa.



Electroforese em gel de agarose

- Separa e analisa DNA
 - Quantifica e Isola uma determinada banda
 - O DNA é visualizado pela adição
 - Brometo de etídio (altamente tóxico e mutagénico, mas dá mais visibilidade)
 - GelRed
 - GelGreen
 - SYBRSafe
- Liga-se ao DNA e fica fluorescente sob luz UV.

Percent Agarose Gel (w/v)	DNA Size Resolution(kb = 1000)
0.5%	1 kb to 30 kb
0.7%	800 bp to 12 kb
1.0%	500 bp to 10 kb
1.2%	400 bp to 7 kb
1.5%	200 bp to 3 kb
2.0%	50 bp to 2 kb

Table 1: Correct Agarose Gel Concentration for Resolving DNA Fragments

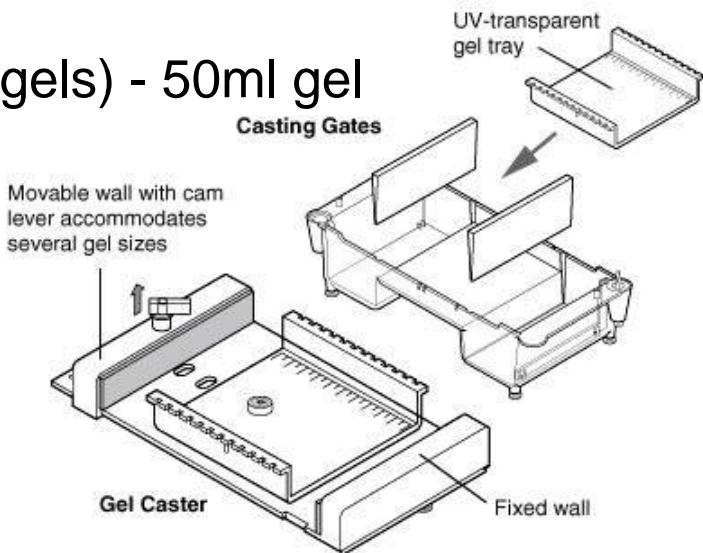
Electroforese em gel de agarose

Polimerização dos géis

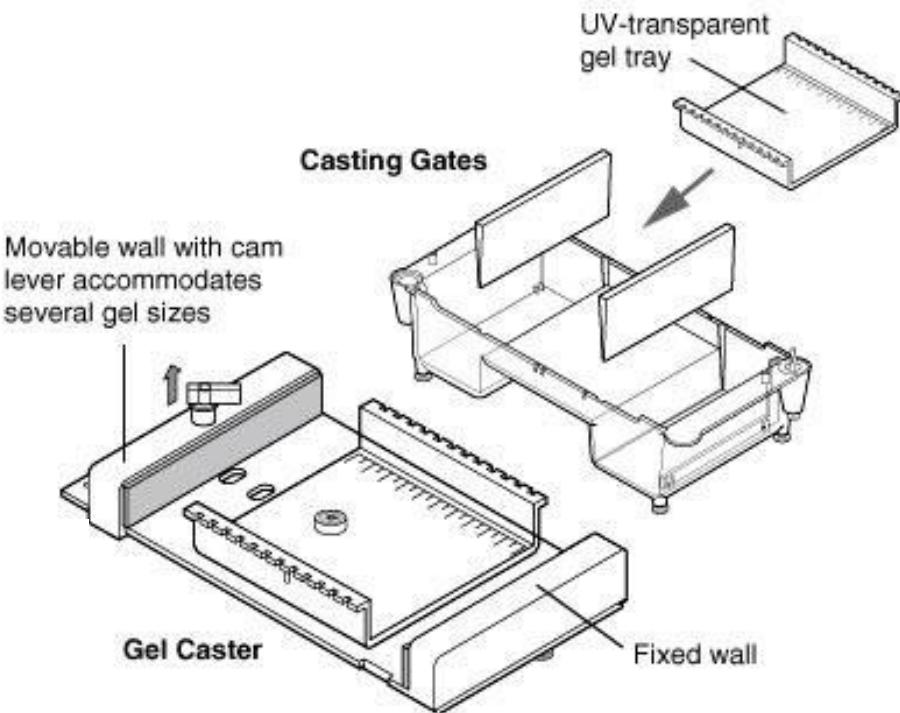
- Há géis que já se compram polimerizados
- Normalmente fazem-se em laboratório

Qual o tanque a escolher

- Número de amostras a carregar
- Velocidade de corrida
- Tinas pequenas 8x10 cm gels (minigels) - 50ml gel
- Tina grande - 100ml gel



Electroforese em gel de agarose



Electroforese em gel de agarose

Qual pente a usar?

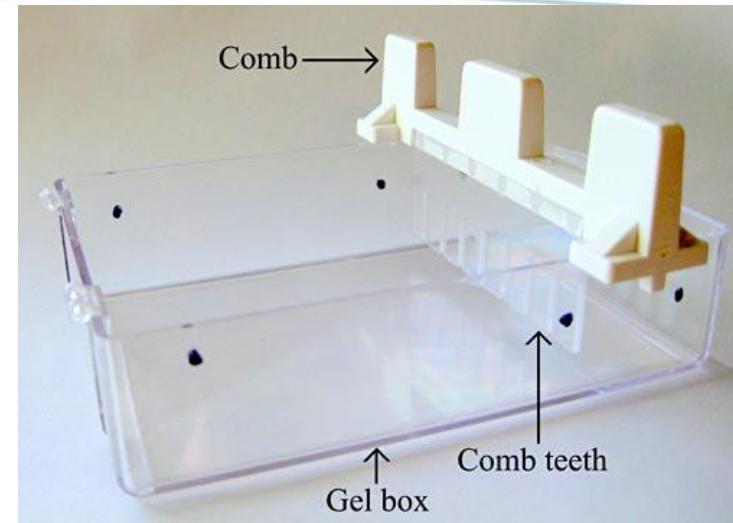
Depende volume que queremos carregar

Volume do poço - Depende largura poço e da espessura do dente do pente, e da espessura do gel

Tina pequena - pente com 8 poços (20 ou 40ul dependendo da espessura do dente) ou 15 poços (20ul)

Tina grande - pente com 15 ou 20 poços

Para além das amostras, não esquecer que temos de carregar, controlo positivo, controlo negativo, geralmente com um poço livre de intervalo e o marcador (1 ou 2 lanes)



Electroforese em gel de agarose

Que quantidade carregar?

- Normalmente, uma banda é visível com 20ng de DNA
- Se queremos apenas confirmar a banda, para mandar sequenciar - 2ul
- Se queremos visualizar a banda num 1º ensaio de PCR- 5ul
- Se queremos cortar a banda para mandar sequenciar - todo o produtos de PCR
(20/25ul se a banda é boa ou 50ul se temos uma banda difusa “faint band”)

Electroforese em gel de agarose

Procedimento

1. Montar o dispositivo para solidificar o gel (pente inclusivé) e verificar que está nivelado

Como fazer o gel de agarose

2. Medir volume TBE (tina pequena-50ml de gel)
 - geis mais finos, dão maior resolução
3. Pesar e adicionar a agarose (1g – se 2%)
4. Dissolver no microondas
 - Cuidado para não queimar as mãos, ferver ou derramar-se
 - Ir agitando ocasionalmente
5. Deixar arrefecer um pouco na bancada
6. Adicionar GreenSafe Premium - 2ul/50ml gel
 - Brometo de etídeo (arrefecer a 60ºC (morno ao toque) para evitar vapores)
7. Agitar sem fazer bolhas

Electroforese em gel de agarose

Procedimento cont.

8. Verter o gel sem fazer bolhas. Eliminar as que possam surgir com uma ponta
9. Esperar 30min que solidifique
10. Mergulhar na tina com tampão de corrida TBE (a mesma que foi usada para fazer o gel), até que a solução cubra superficialmente os poços
 - Atenção, os poços devem estar junto do pólo negativo (PP-Preto-Poços)
11. Retirar com cuidado o pente
12. Se não quiser usar imediatamente, tapar o gel com papel de alumínio para evitar a degradação do Greensafe com a luz
13. Carregar o gel (1º amostras, 2º Controlo negativo com 1 poço de intervalo, 3º marcador (no 1º e último poço))
14. Colocar a correr (gel pequeno)
 - Voltagem pequeno :80V, 30min
 - Voltagem grande: 100V, 45min
 - Corridas maiores, separam melhor os fragmentos

Electroforese em gel de agarose

Soluções Tampão

- **TBE** (Tris/Borate/EDTA)
 - solução stock a 5X, ou outra concentração mais elevada, mas é menos estável pois há precipitação durante o armazenamento
 - solução ao uso 1X
- Tris-faz com que o DNA continue solúvel em água
- EDTA - quelante de iões divalentes, protege o DNA da degradação enzimática (iões são cofatores de muitas enzimas)
- Borato-inibidor de enzimas

- **TAE** (Tris/Acetate/EDTA)
 - Menor capacidade tampão (tem de se reutilizar menos vezes), mas DNA corre mais depressa. Necessita menos procedimentos para extrair DNA do gel
 - Porém há quem use os 2 tampões da mesma maneira

Electroforese em gel de agarose

Recipes for TAE and TBE Electrophoresis Buffers

Agarose gels are generally run two types of electrophoresis buffers. Nucleic acid agarose gel electrophoresis is usually conducted with either Tris-acetate-EDTA (TAE) buffer or Tris-borate-EDTA (TBE) buffer. While TAE buffer provides faster electrophoretic migration of linear DNA and better resolution of supercoiled DNA, TBE buffers have a stronger buffering capacity for longer or higher voltage electrophoresis runs.

Note: Because of higher voltages and resulting higher currents often used with Sub-Cell® Model 96 and 192 cells, we recommend that only TBE buffers be used for electrophoresis with these specific systems.

Tris-acetate-EDTA (TAE) buffer

TAE is often prepared in concentrated stock solutions of 10x or 50x in the laboratory. A 1x working solution is prepared prior to electrophoresis.

Composition of 1x TAE buffer

- 40 mM Tris (pH 7.6)
- 20 mM acetic acid
- 1 mM EDTA

Preparation of 50x TAE stock solution

To prepare 1 liter of 50x TAE dissolve following components in 600 ml of deionized water:

- 242 g Tris base (FW = 121)
- 57.1 ml glacial acetic acid
- 100 ml 0.5 M EDTA (pH 8.0)

Adjust the final volume to 1 liter with deionized water.

To prepare a 1x working solution from 50x stock buffer mix 50x stock buffer with DNase free deionized water at 1:4 ratio.

Tris-borate-EDTA (TBE) buffer

TBE buffer can be made and stored in concentrated stocks of 5x or 10x.

Composition of 1x TBE buffer

- 89 mM Tris (pH 7.6)
- 89 mM boric acid
- 2 mM EDTA

Preparation of 10x TBE stock solution

To prepare 1 liter of 10x TBE dissolve following components in 600 ml deionized water:

- 108 g Tris base (FW = 121)
- 55 g boric acid (FW = 61.8)
- 40 ml 0.5 M EDTA (pH 8.0)

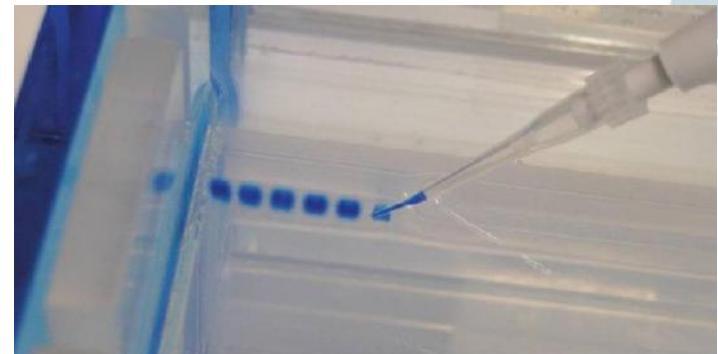
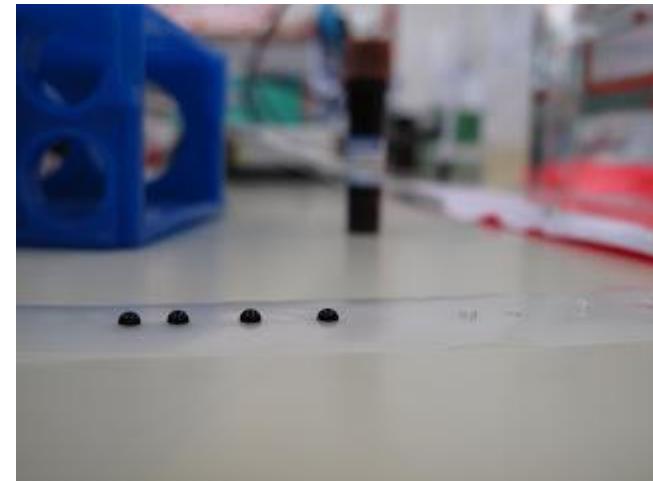
Adjust final volume to 1 liter with deionized water.

To prepare a 1x working solution from 10x stock buffer, mix 10x stock buffer with DNase free deionized water at 1:9 ratio.

Electroforese em gel de agarose

Como carregar as amostras

1. Spin down da amostras a carregar
2. Cortar um pedaço de Parafilm®
3. Decidir o volume a carregar
4. Colocar cerca de 0,2% Loading Buffer no Parafilm
 - ex: 2ul por 10ul de amostra
 - nosso caso 1ul por 5ml produto PCR ou DNA genómico
4. Pipetar a amostra, despejar a amostra na gota de loading dye, agitar cima e baixo, sem bolhas e carregar no poço



Nota: Algumas mix de PCR já incorporam o loading dye

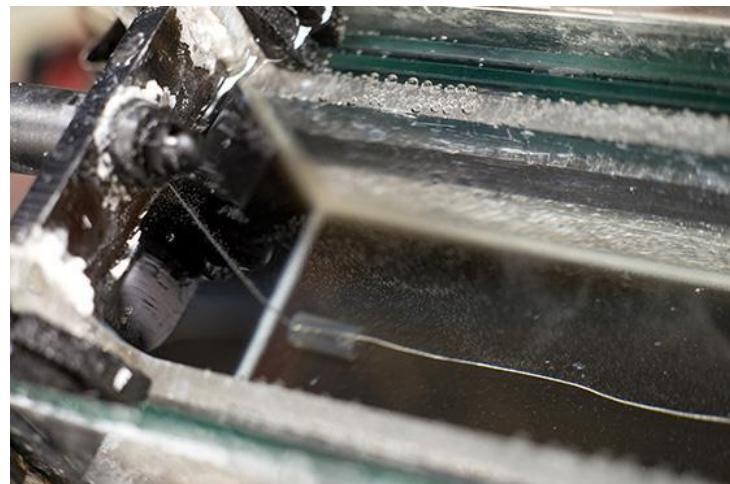
Electroforese em gel de agarose

5. Carregar 2ul de Marcador/Ruller/Ladder

- No inicio e fim das amostras
 - Se for possível, deixar os poços dos extremos pois costumam correr pior
 - Resolução do marcador escolhida relativamente ao peso do meu fragmento de interesse
-
- Carregar da esquerda para a direita, com os poços em cima (no gel). Por convenção, publicam-se os resultados assim.

6. Fechar a tampa, ligar aparelho, colocar a correr (80V, 30min)

7. Verificar se está a passar corrente eléctrica (ver bolhinhas à volta do fio)



Electroforese em gel de agarose

Marcadores

Alto peso molecular

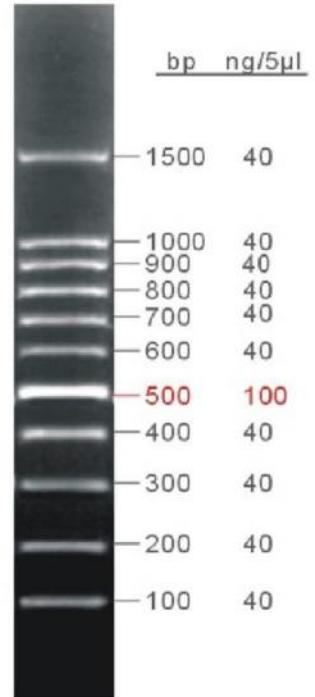
NzyDNA ladder III



Band size (bp)	ng/band
10000	100
7500	75
6000	60
5000	50
4000	40
3000	30
2500	25
2000	20
1400	14
1000	100
800	80
600	60
400	40
200	20

Baixo peso molecular

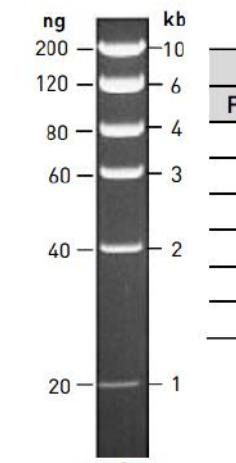
Grisp ladder 100pb



1.7% Agarose | 0.5x TBE | 50V - 1h
minigel - 5µL/lane

Marcador quantitativo

Mass Ladder Thermofisher



1% agarose gel stained with ethidium bromide
4 µL/lane

Fragment size	Amount of DNA (ng) in each band		
	2 µL	4 µL	8 µL
10,000 bp	100 ng	200 ng	400 ng
6000 bp	60 ng	120 ng	240 ng
4000 bp	40 ng	80 ng	160 ng
3000 bp	30 ng	60 ng	120 ng
2000 bp	20 ng	40 ng	80 ng
1000 bp	10 ng	20 ng	40 ng

Electroforese em gel de agarose

Loading buffers

- Dá cor à amostra
 - bromophenol blue
 - xylene cyanol
 - Orange G
- Dá densidade à amostra
 - Glicerol
 - sucrose
- Permite monitorizar a corrida
- Facilita o carregamento nos poços
- Carregados negativamente, logo migram no mesmo sentido que o DNA

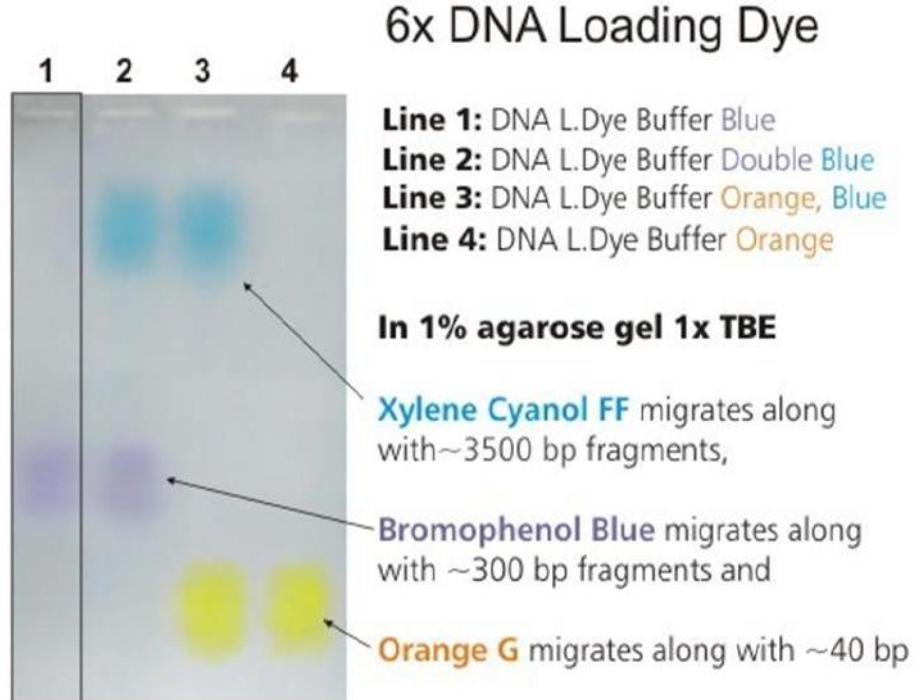


Eletroforese em gel de agarose

Loading buffers

Typical recipe

- 25 mg bromophenol blue or xylene cyanol
- 4 g sucrose
- H₂O to 10 mL
- The exact amount of dye is not important
- Store at 4°C to avoid mould growing in the sucrose. 10 mL of loading buffer will last for years.
- Escolher o corante com base no fragmento que queremos ver, para não impedir a visualização da banda



Visualização das amostras

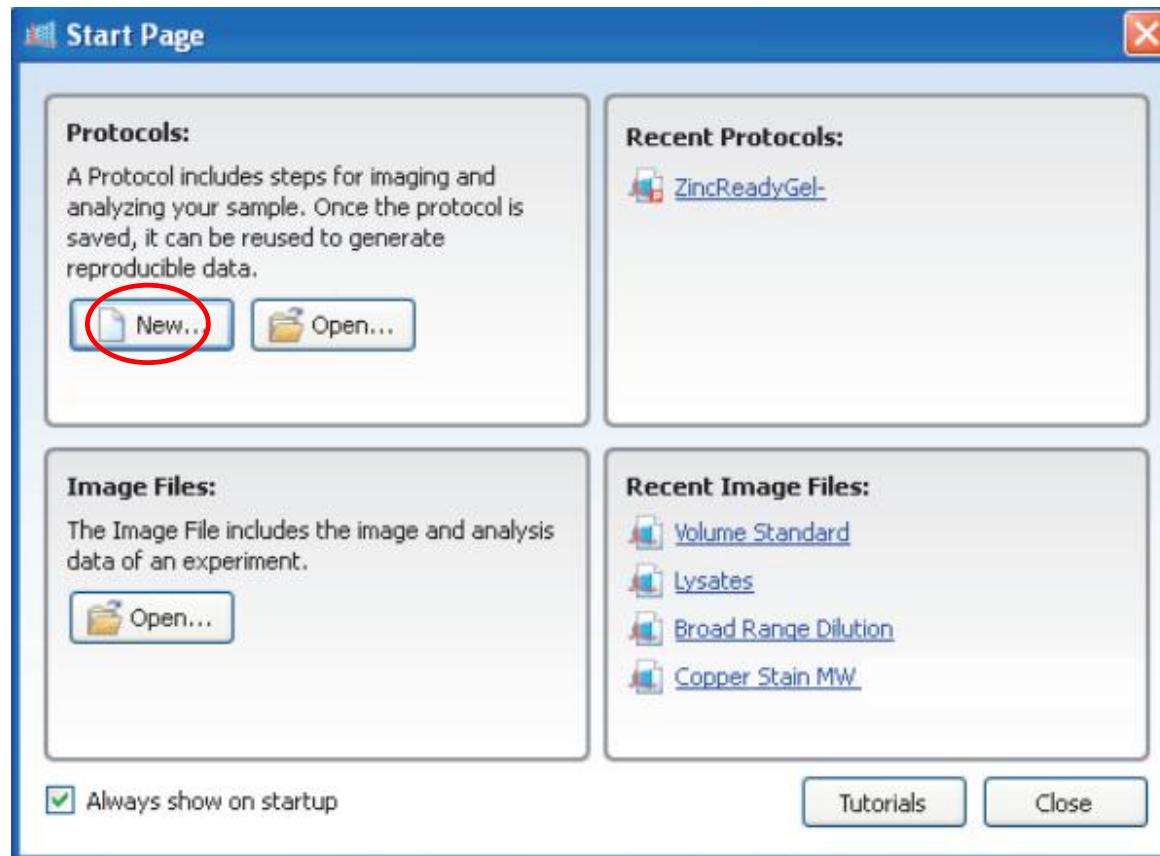
1. Após corrida do gel, desligar a Tina de Eletroforese
2. Colocar o gel no tabuleiro com papel absorvente, para retirar o excesso de solução tampão
3. Colocá-lo (com o suporte em acrílico) no transiluminador
-se o seu suporte não for apropriado para ver directamente o gel, retirá-lo do suporte
4. Ligar o transiluminador e o PC



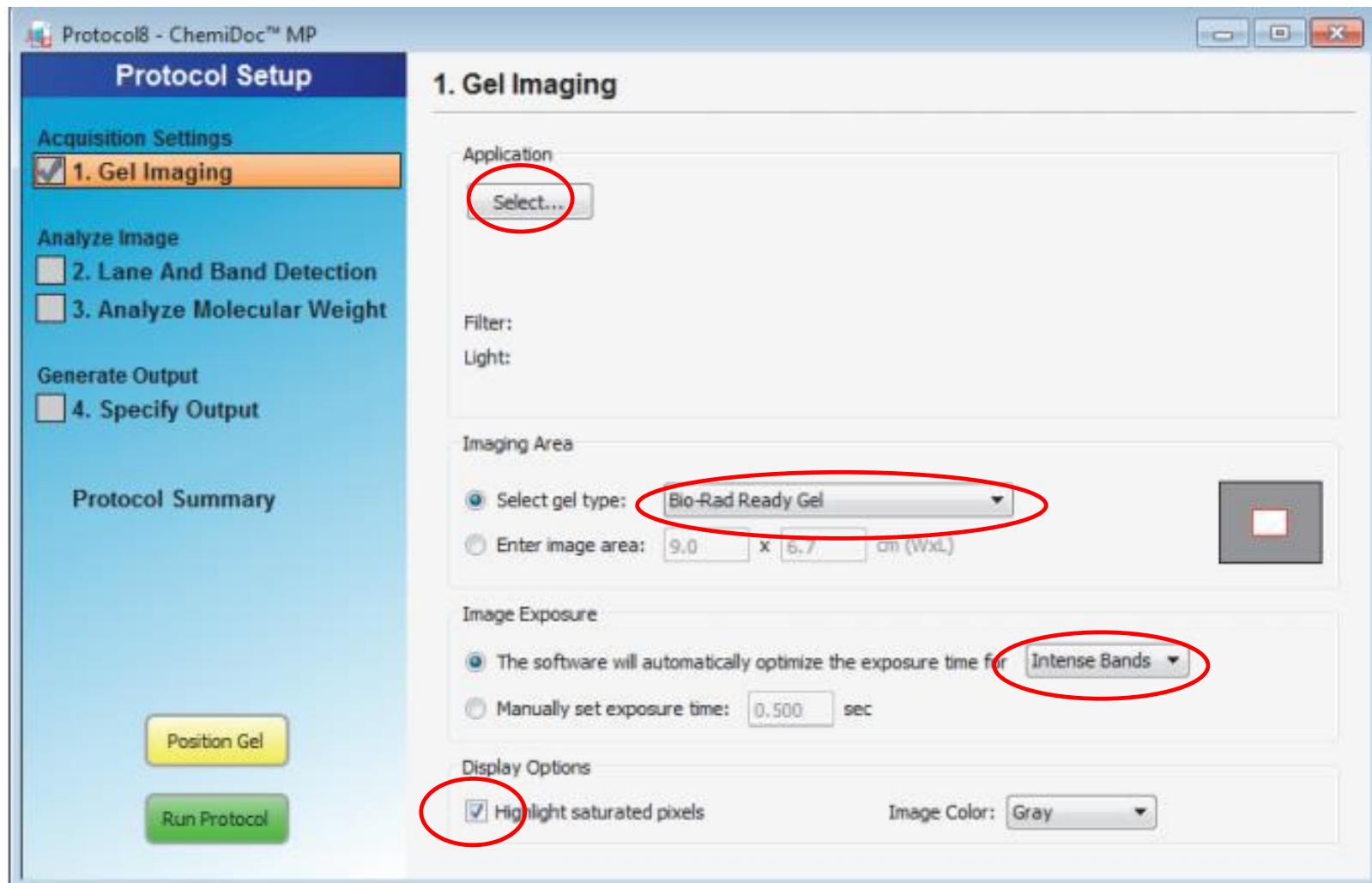
Os UV são carcinogénicos- proteger pele e olhos.

Visualização das amostras

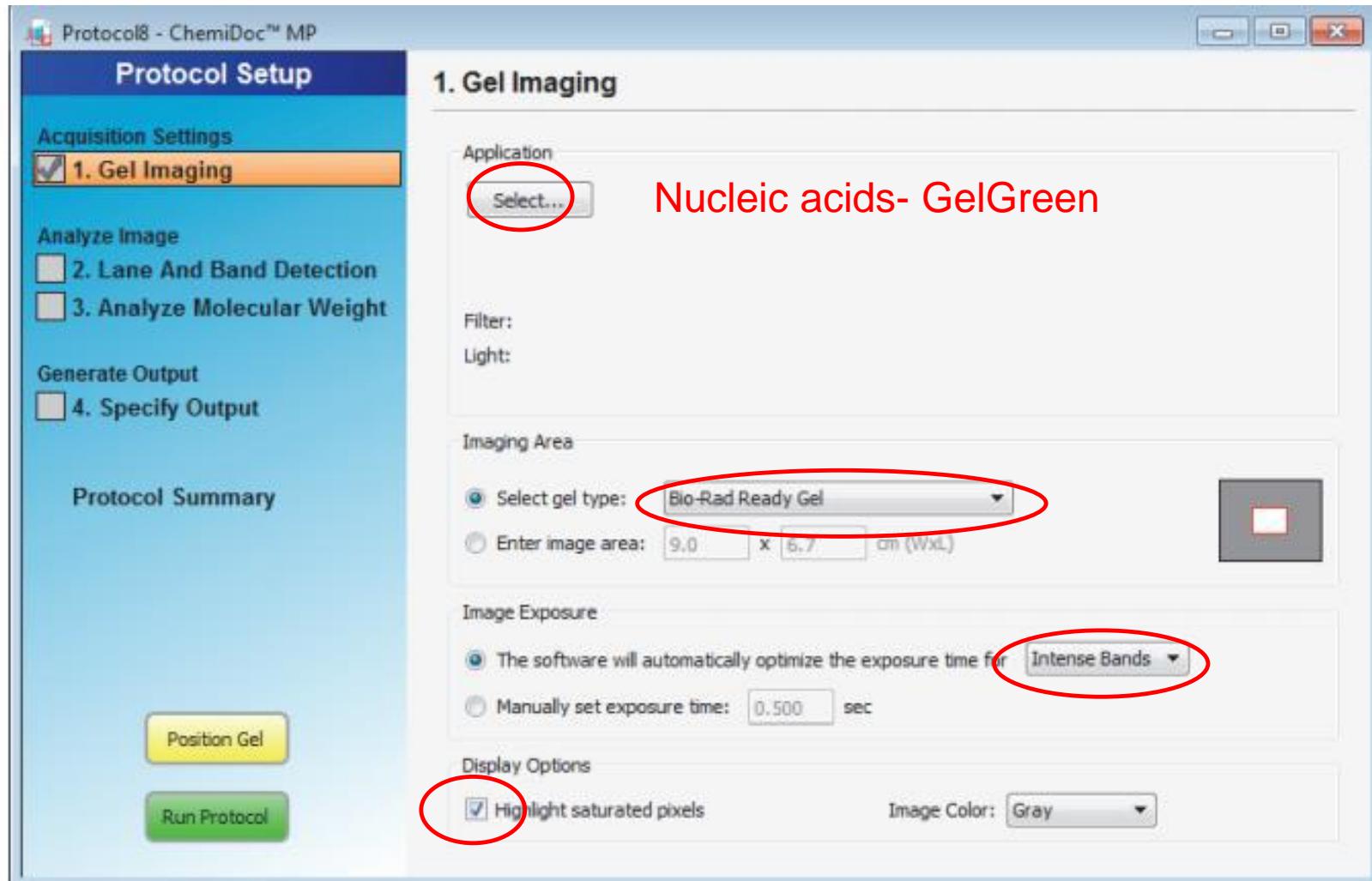
1. Abrir atalho gel-doc



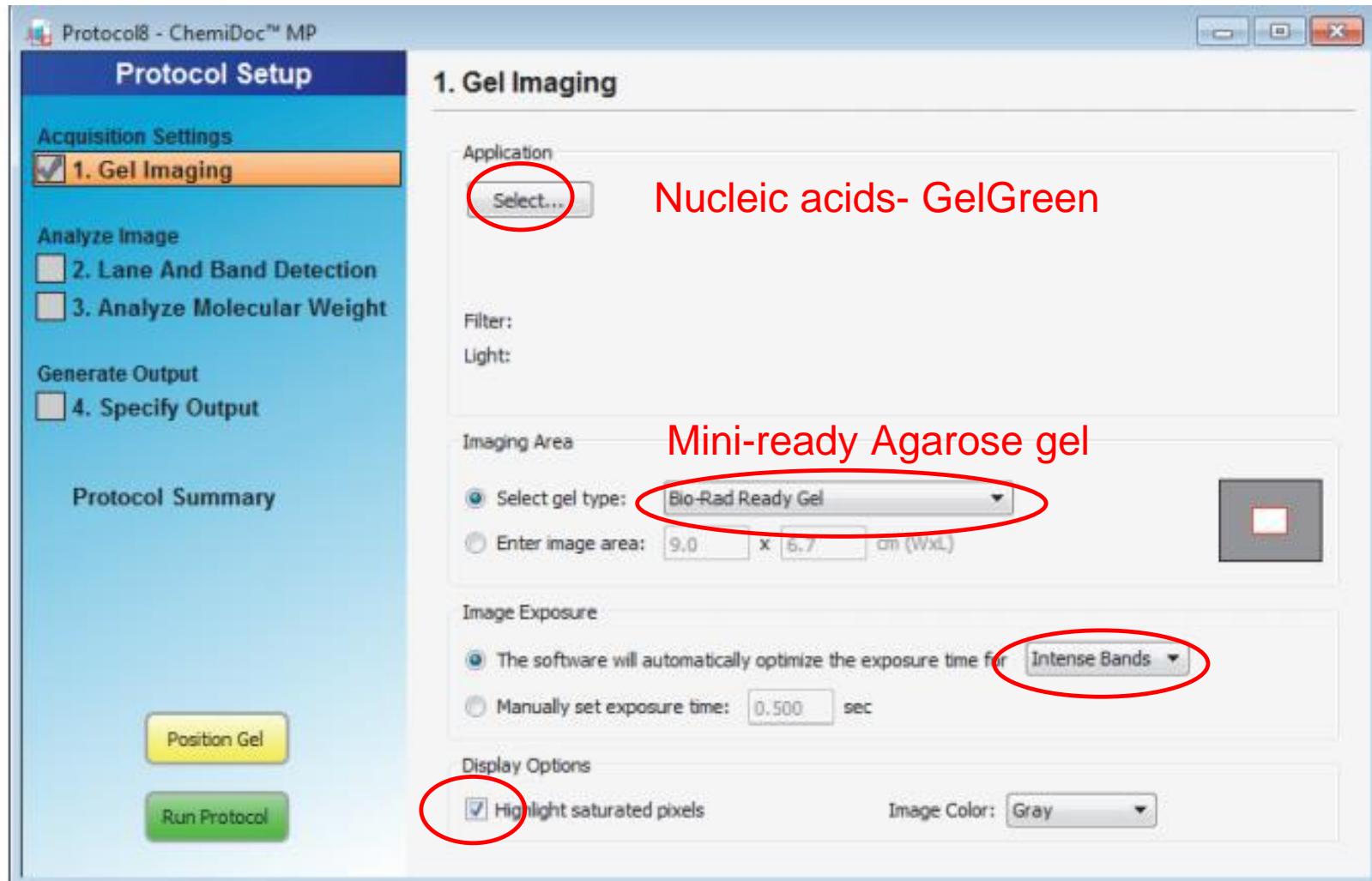
Eletroforese em gel de agarose



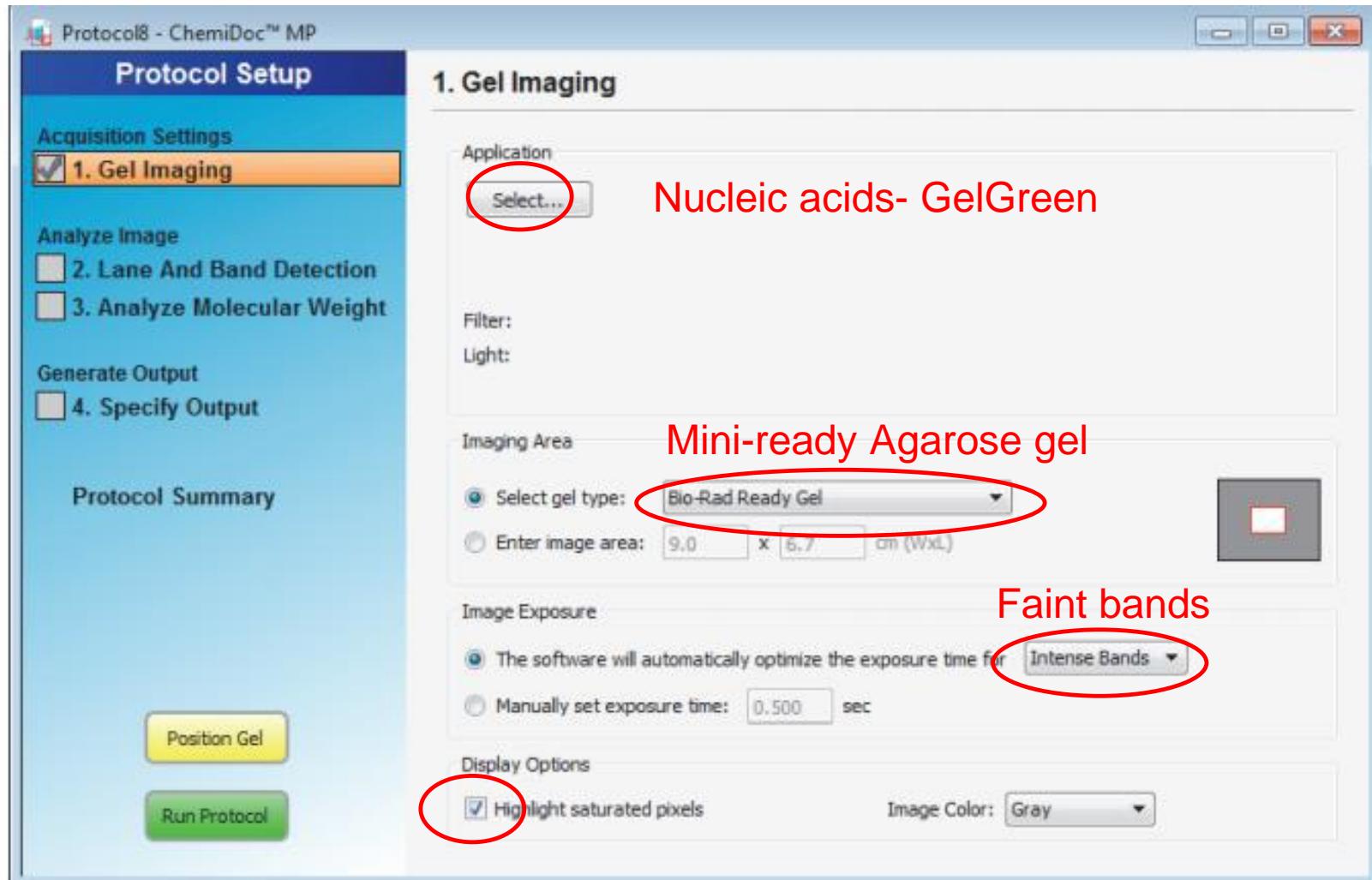
Eletroforese em gel de agarose



Eletroforese em gel de agarose



Eletroforese em gel de agarose



Nucleic acids- GelGreen

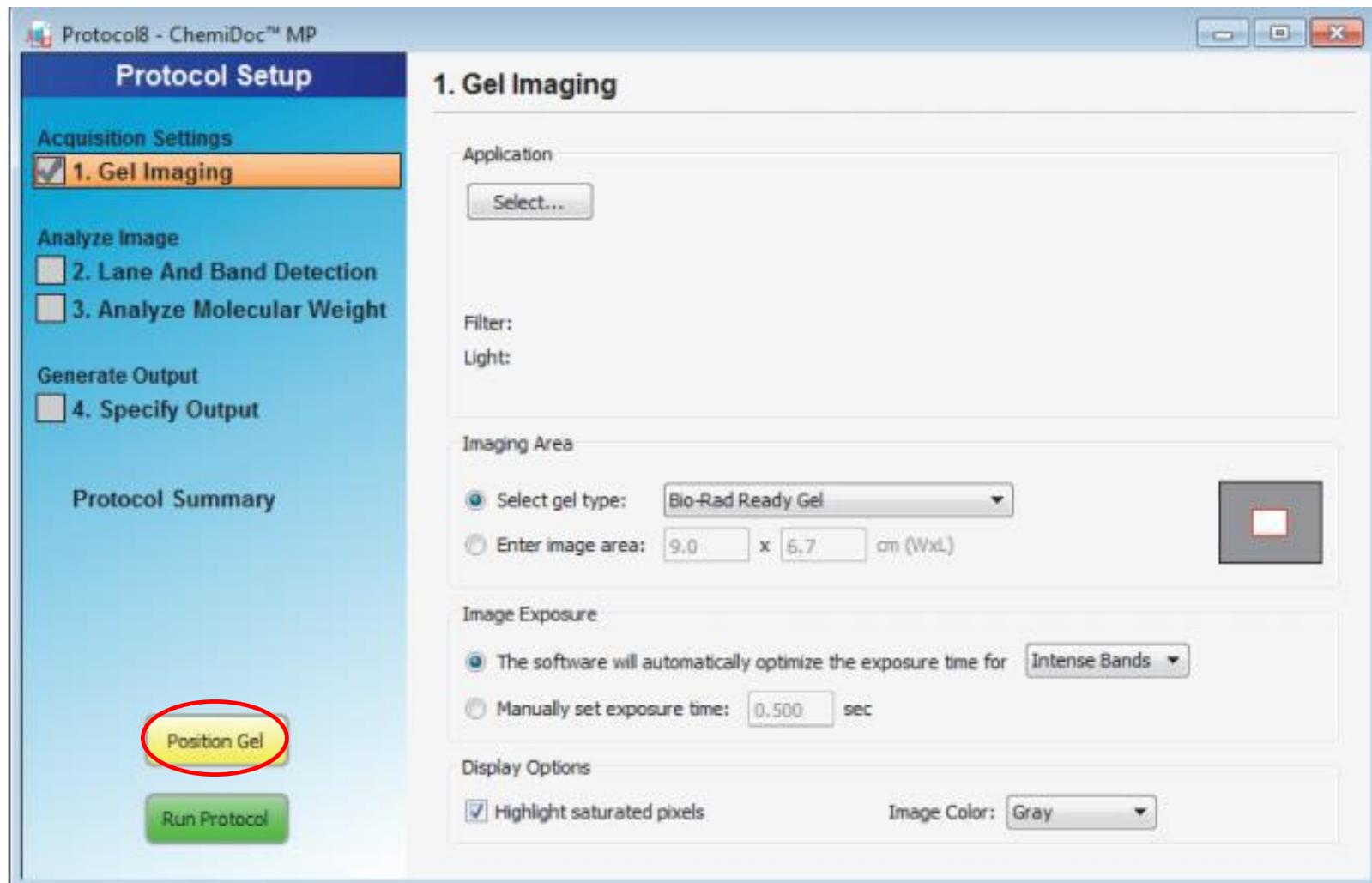
Mini-ready Agarose gel

Faint bands

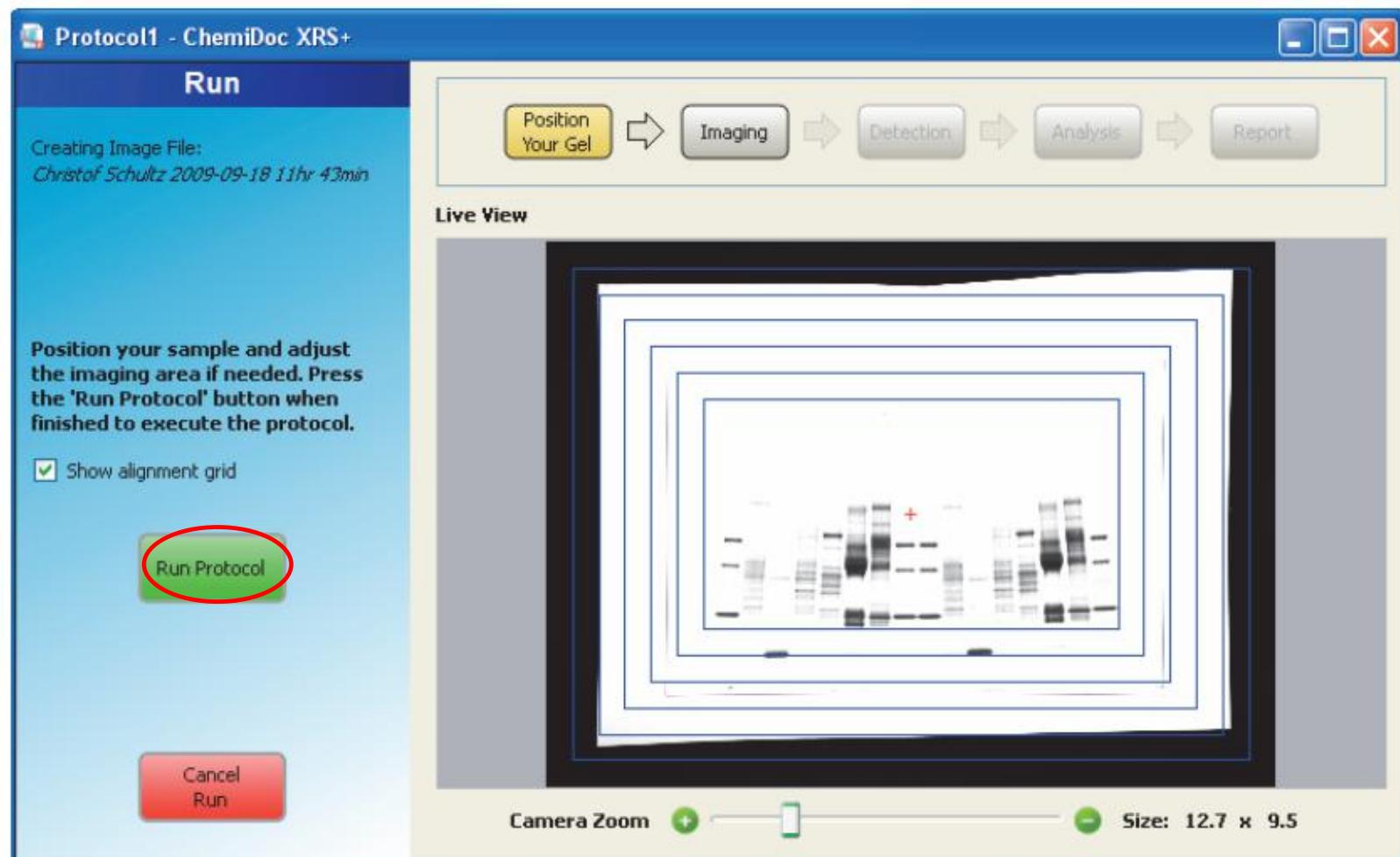
Intense Bands

Gray

Eletroforese em gel de agarose



Eletroforese em gel de agarose



Quantificação do DNA

Quantificação de DNA genómico em gel de agarose

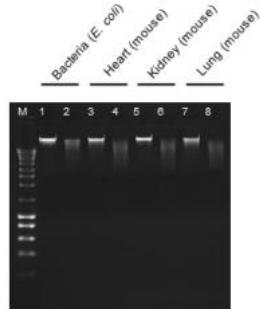
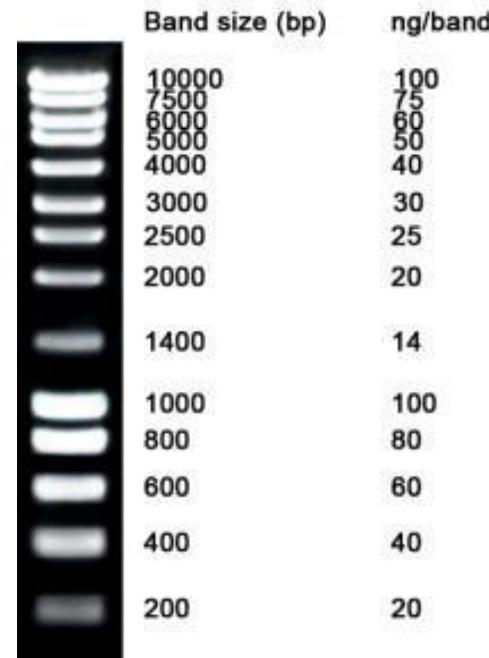
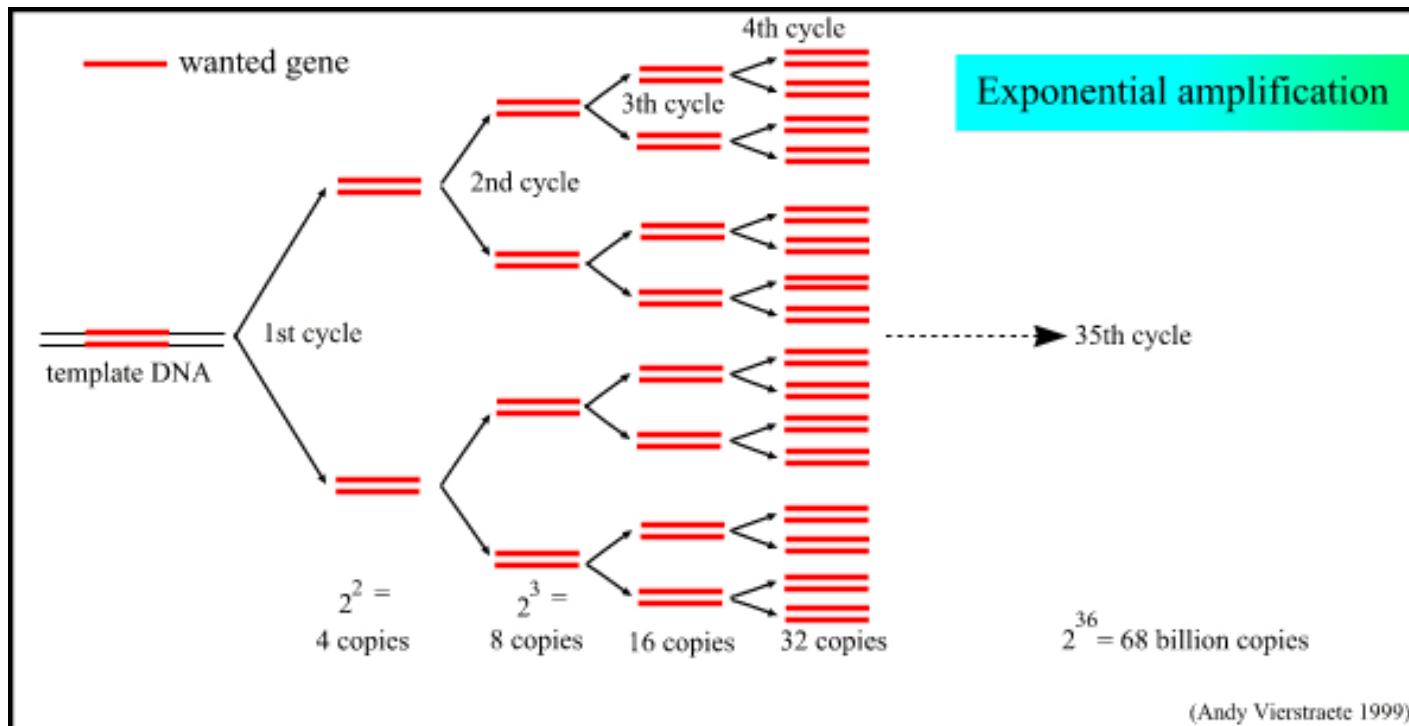


Fig. 1. Genomic DNA from various sources isolated with NZY Tissue gDNA Isolation kit. 0.25 µg of each isolated DNA was analysed on a 0.8% agarose gel. Lanes 2, 4, 6 and 8: gDNA digested with *EcoR* I. Lanes 1, 3, 5 and 7: undigested gDNA. M: NZYDNA Ladder III (MB0440); 1, 2: gDNA from bacteria (*E. coli*); 3, 4: gDNA from mouse heart; 5, 6: gDNA from mouse kidney; 7, 8: gDNA from mouse lung.



- Descrito por Kary Mullis, 1983,
 - Prémio Nobel da Química
- É um método de amplificação (de criação de múltiplas cópias) de DNA
- Aplicações
 - Investigação médica e biológica
 - detecção de doenças hereditárias,
 - testes de paternidade,
 - exames para detecção de agentes patogénicos
 - etc

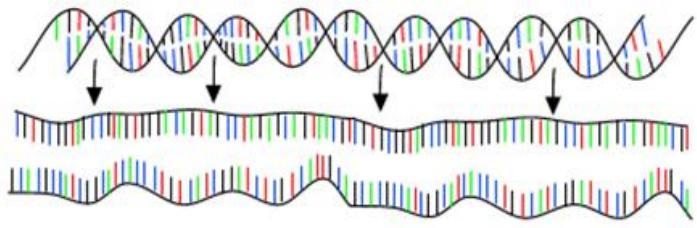
PCR-Polymerase Chain reaction



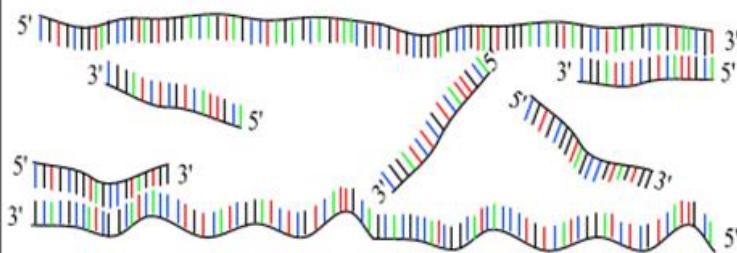
PCR-Polymerase Chain reaction

PCR : Polymerase Chain Reaction

30 - 40 cycles of 3 steps :

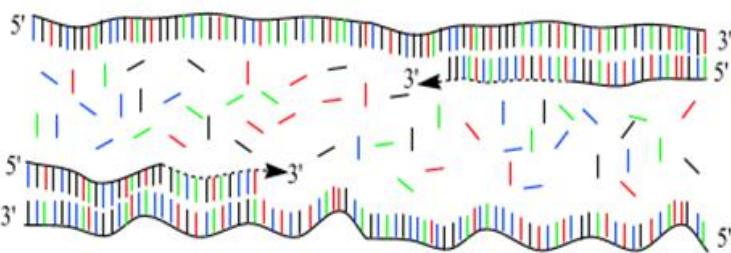


1 minut 94 °C



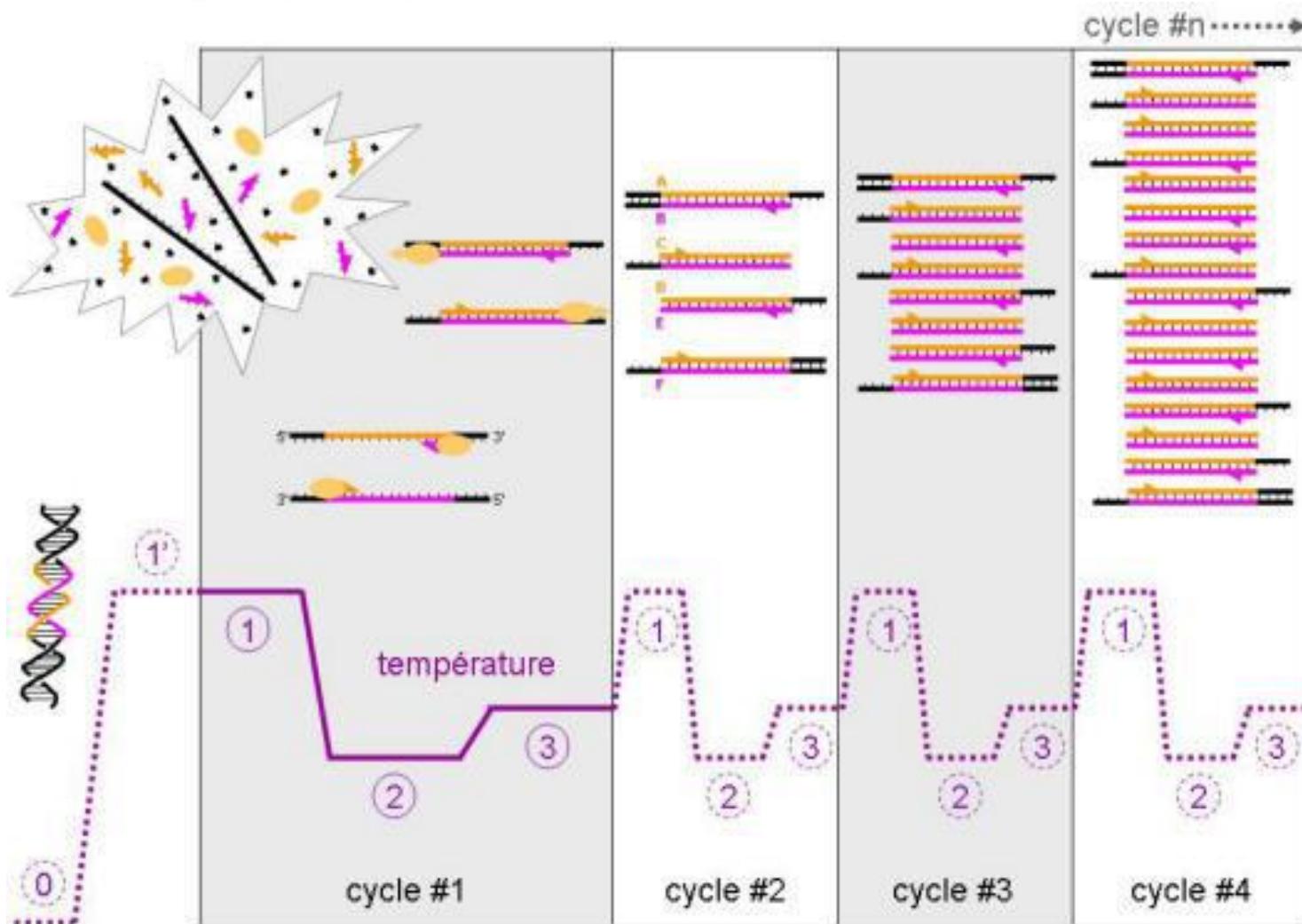
45 seconds 54 °C

forward and reverse
primers !!!



2 minutes 72 °C
only dNTP's

(Andy Vierstraete 1999)



PCR-Polymerase Chain reaction

Componentes:

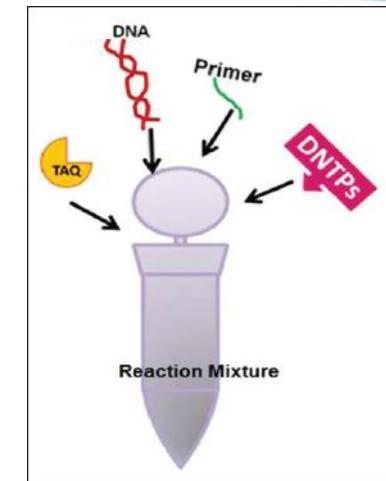
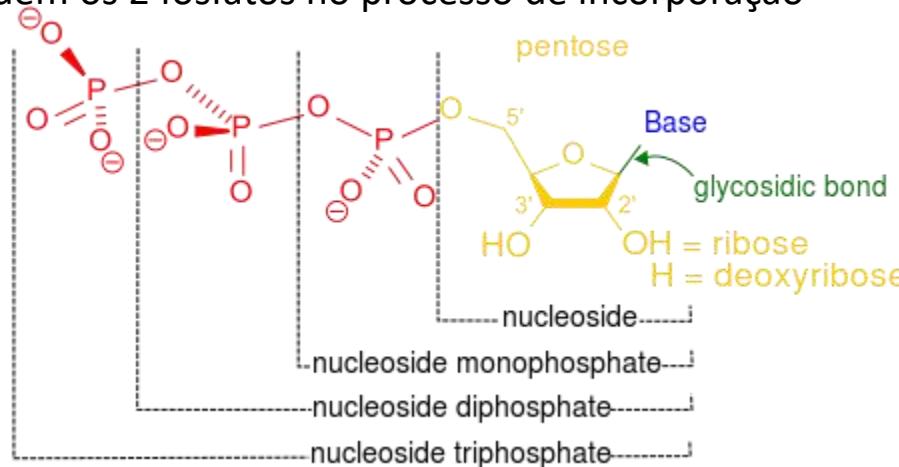
1. Cadeia molde “template” de DNA

10 a 100ng/ul

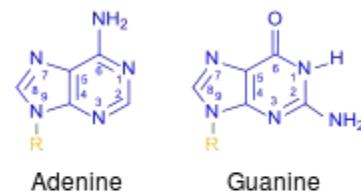
Diluir se a concentração for elevada, pois elevadas concentrações inibem reação

2. dNTPs (desoxirribonucleotídeos trifosfatos), que são as bases nitrogenadas ligadas 3 três fosfatos (dATP, dGTP, dTTP, dCTP)

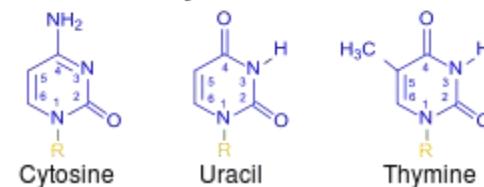
Perdem os 2 fosfatos no processo de incorporação



Purines



Pyrimidines



dNTPs Working solution

Exercício:

- Temos dATP, dCTP, dGTP, dTTP – 100mM each
- Queremos 0,2mM each numa reação de PCR com Volume total 20ul
- Como fazemos nossa solução de trabalho?

dNTPs Working solution

Exercício:

- Temos dATP, dCTP, dGTP, dTTP – 100mM each
- Queremos 0,2mM each numa reação de PCR com Volume total 20ul
- Como fazemos nossa solução de trabalho?

Como fazer:

$$C_i \times V_i = C_f \times V_f$$

$$C_i (\text{working solution}) \times 1\text{ul} = 0,2\text{mM} (\text{no PCR}) \times 20\text{ul} (\text{PCR})$$

$$C_i = 4\text{mM}$$

dNTPs Working solution

Exercício:

- Temos dATP, dCTP, dGTP, dTTP – 100mM each
- Queremos 0,2mM each numa reação de PCR com Volume total 20ul
- Como fazemos nossa solução de trabalho de 100ul?

Como fazer:

$$C_i \times V_i = C_f \times V_f$$

$$C_i (\text{working solution}) \times 1\text{ul} = 0,2\text{mM} (\text{no PCR}) \times 20\text{ul} (\text{PCR})$$

$$C_i = 4\text{mM}$$

$$C_i \times V_i = C_f \times V_f$$

$$100\text{mM} \times V_i (\text{compra}) = 4\text{mM} \times 100\text{ul} (\text{normalmente} - \text{volume working solution})$$

$$V_i = 4\text{ul}$$

dNTPs Working solution

Exercício:

- Temos dATP, dCTP, dGTP, dTTP – 100mM each
- Queremos 0,2mM each numa reação de PCR com Volume total 20ul
- Como fazemos nossa solução de trabalho?

Como fazer:

dATP (100mM)- 4ul

dCTP (100mM)- 4ul

dGTP (100mM)- 4ul

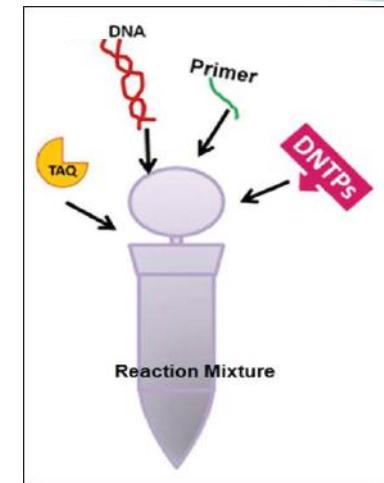
dTTP (100mM)- 4ul

H₂O – 84ul

Mix e spin

Componentes:

3. Água de uso molecular
4. Primers (também chamados de oligonucleotídeos ou iniciadores)
 - Pequena sequencia complementar ao DNA template, que flanqueia a minha região de interesse
 - Compram-se em solução ou liofilizados
 - Reconstituem-se em água de uso molecular
 - Manipulam-se numa câmara de PCR
 - Convém usar pontas com filtro



Como Reconstituir e fazer working solution dos Primers? D1R/D3Ca

**Invitrogen Custom Primers
Certificate of Analysis**

Primer 4: P1/

Primer Name: PnAIR ITS1
 Researcher: Alexandra Silva
 Sequence (5' to 3'): CTT TAG GTC ATT TGG TT
 Molecular Weight (µg/µmole): 5198.4
 Micromolar Extinction Coeff(OD/µmol): 176.3
 Purity Desalting
 Tm (1 M Na⁺) 56
 Tm (50 mM Na⁺) 35
 % GC 35
 Notes:

Primer Number:	M7303B04 (B04)
Primer Length:	17
Scale of Synthesis:	50n mol
µg per OD:	29.5
nmoles per OD:	5.7
OD's	7.30
µg's*	215.25
nmoles	41.4 (41.4)
Coupling Eff.	99%

Primer 5: PN / UNIversal

Primer Name: D1R
 Researcher: Alexandra Silva
 Sequence (5' to 3'): ACC CGC TGA ATT TAA GCA TA
 Molecular Weight (µg/µmole): 6086.0
 Micromolar Extinction Coeff(OD/µmol): 226.0
 Purity Desalting
 Tm (1 M Na⁺) 64
 Tm (50 mM Na⁺) 43
 % GC 40
 Notes:

Primer Number:	M7303B05 (B05)
Primer Length:	20
Scale of Synthesis:	50n mol
µg per OD:	26.9
nmoles per OD:	4.4
OD's	11.10
µg's*	298.81
nmoles	49.1 (49.1)
Coupling Eff.	99%

Primer 6: PN

Primer Name: D2C
 Researcher: Alexandra Silva
 Sequence (5' to 3'): CCT TGG TCC GTG TTT CAA GA
 Molecular Weight (µg/µmole): 6100.0
 Micromolar Extinction Coeff(OD/µmol): 207.0
 Purity Desalting
 Tm (1 M Na⁺) 68
 Tm (50 mM Na⁺) 47
 % GC 50
 Notes:

Primer Number:	M7303B06 (B06)
Primer Length:	20
Scale of Synthesis:	50n mol
µg per OD:	29.5
nmoles per OD:	4.8
OD's	12.20
µg's*	359.52
nmoles	58.9
Coupling Eff.	97%

10 H₂O
 + 10 P1(DNA) (10x)

FOR LABORATORY RESEARCH USE ONLY.

CAUTION: Not for diagnostic use. The safety and efficacy of this product in diagnostic or other clinical uses has not been established.

invitrogen
by life technologies®

Using the nanomole quantity - to reconstitute to a given concentration, convert the n mole figure to umole, and then divide the desired concentration in umole/litre. For example, to make a 100 umole primer stock solution, assuming 24nmole yield
 $24\text{nmole} \times 1\text{umole}/1000\text{nmole} = 0.024\text{ umole}$
 $0.024\text{umole}/100\text{umole/litre} = 0.00024\text{ L}$
 $0.00024\text{ L} \times 1000\text{mL/L} = 0.24\text{ml or } 240\mu\text{l}$

* Other supporting information available on-line.

Como Reconstituir e
fazer working solution dos Primers?
D1R/D3Ca

Solução stock-100uM
VH2O (uL)= nmoles x 10

Exercício?

Working solution 10uM, 100uL?

Invitrogen Custom Primers Certificate of Analysis		ALFAGENE LDA	
		Order Number: 279691 79	
		Order Date: 05/12/13	
Primer 4: P1/		Primer Number: M7303B04 (B04)	
Primer Name: PnAIR ITS1		Primer Length: 17	
Researcher: Alexandra Silva		Scale of Synthesis: 50n mol	
Sequence (5' to 3') CTT TAG GTC ATT TGG TT		µg per OD:	29.5
Molecular Weight (µg/µmole): 5198.4		nmoles per OD:	5.7
Micromolar Extinction Coeff(OD/µmol): 176.3		OD's	7.30
Purity	Desalting	µg's*	215.25
Tm (1 M Na ⁺)	56	nmoles	41.4 (1.4)
Tm (50 mM Na ⁺)	35	Coupling Eff.	99%
% GC	35		
Notes:			
Primer 5: PN / Univeral		Primer Number: M7303B05 (B05)	
Primer Name: D1R		Primer Length: 20	
Researcher: Alexandra Silva		Scale of Synthesis: 50n mol	
Sequence (5' to 3') ACC CGC TGA ATT TAA GCA TA		µg per OD:	26.9
Molecular Weight (µg/µmole): 6086.0		nmoles per OD:	4.4
Micromolar Extinction Coeff(OD/µmol): 226.0		OD's	11.10
Purity	Desalting	µg's*	298.91
Tm (1 M Na ⁺)	64	nmoles	49.1 (4.91)
Tm (50 mM Na ⁺)	43	Coupling Eff.	99%
% GC	40		
Notes:			
Primer 6: P2		Primer Number: M7303B06 (B06)	
Primer Name: D2C		Primer Length: 20	
Researcher: Alexandra Silva		Scale of Synthesis: 50n mol	
Sequence (5' to 3') CCT TGG TCC GTG TTT CAA GA		µg per OD:	29.5
Molecular Weight (µg/µmole): 6100.0		nmoles per OD:	4.8
Micromolar Extinction Coeff(OD/µmol): 207.0		OD's	12.20
Purity	Desalting	µg's*	359.52
Tm (1 M Na ⁺)	68	nmoles	58.9
Tm (50 mM Na ⁺)	47	Coupling Eff.	97%
% GC	50		
Notes:			
$10 \text{ H}_2\text{O}$ $+ 10 \text{ P}_1(\text{mM}) \rightarrow \text{(molar)}$			
FOR LABORATORY RESEARCH USE ONLY. CAUTION: Not for diagnostic use. The safety and efficacy of this product in diagnostic or other clinical uses has not been established.			
Using the nanomole quantity - to reconstitute to a given concentration, convert the nmole figure to umole, and then divide the desired concentration in umole/litre. For example, to make a 100 umole primer stock solution, assuming 24nmole yield $24\text{nmole} \times 1\text{umole}/1000\text{nmole} = 0.024\text{ umole}$ $0.024\text{umole}/100\text{umole/litre} = 0.00024\text{ L}$ $0.00024\text{ L} \times 1000\text{mL/L} = 0.24\text{ml or } 240\mu\text{l}$			
<small>* Other supporting information available on-line.</small>			

invitrogen
by life technologies®

Como Reconstituir e fazer working solution dos Primers? D1R/D3Ca

Solução stock-100uM
 VH₂O (uL)= nmoles x 10

Exercício?

Working solution 10uM, 100uL?

$$C_i \times V_i = C_f \times V_f$$

$$100\text{uM} \times V_i = 10\text{uM} \times 100\text{uL}$$

$$V_i = 10 \text{ uL}$$

$$\text{Working solution } 10\text{uM} = 10\text{uL stock} + 90 \text{ uL H}_2\text{O}$$

Invitrogen Custom Primers Certificate of Analysis

Primer 4: P1/

Primer Name: PnAIR ITS1
 Researcher: Alexandra Silva
 Sequence (5' to 3'): CTT TAG GTC ATT TGG TT
 Molecular Weight (µg/µmole): 5198.4
 Micromolar Extinction Coeff(OD/µmol): 176.3
 Purity Desalting
 Tm (1 M Na⁺) 56
 Tm (50 mM Na⁺) 35
 % GC 35
 Notes:

Primer Number:	M7303B04 (B04)
Primer Length:	17
Scale of Synthesis:	50n mol
µg per OD:	29.5
nmoles per OD:	5.7
OD's	7.30
µg's*	215.25
nmoles	41.4 (41.4)
Coupling Eff.	99%

Primer 5: PN/UNIVERSAL

Primer Name: D1R
 Researcher: Alexandra Silva
 Sequence (5' to 3'): ACC CGC TGA ATT TAA GCA TA
 Molecular Weight (µg/µmole): 6086.0
 Micromolar Extinction Coeff(OD/µmol): 226.0
 Purity Desalting
 Tm (1 M Na⁺) 64
 Tm (50 mM Na⁺) 43
 % GC 40
 Notes:

Primer Number:	M7303B05 (B05)
Primer Length:	20
Scale of Synthesis:	50n mol
µg per OD:	26.9
nmoles per OD:	4.4
OD's	11.10
µg's*	298.91
nmoles	49.1 (49.1)
Coupling Eff.	99%

Primer 6: P2N

Primer Name: D2C
 Researcher: Alexandra Silva
 Sequence (5' to 3'): CCT TGG TCC GTG TTT CAA GA
 Molecular Weight (µg/µmole): 6100.0
 Micromolar Extinction Coeff(OD/µmol): 207.0
 Purity Desalting
 Tm (1 M Na⁺) 68
 Tm (50 mM Na⁺) 47
 % GC 50
 Notes:

Primer Number:	M7303B06 (B06)
Primer Length:	20
Scale of Synthesis:	50n mol
µg per OD:	29.5
nmoles per OD:	4.8
OD's	12.20
µg's*	359.52
nmoles	58.9
Coupling Eff.	97%

*10 H₂O
 + 10 P1 (100uL) → (100uL)*

FOR LABORATORY RESEARCH USE ONLY.

CAUTION: Not for diagnostic use. The safety and efficacy of this product in diagnostic or other clinical uses has not been established.

invitrogen
by life technologies®

Using the nanomole quantity - to reconstitute to a given concentration, convert the nmole figure to umole, and then divide the desired concentration in umole/litre. For example, to make a 100 umole primer stock solution, assuming 24nmole yield
 24nmole x 1umole/1000nmole = 0.024 umole
 0.024umole/100umole/litre = 0.00024 L
 0.00024 L x 1000mL = 0.24ml or 240ul

* Other supporting information available on-line.

PCR-Polymerase Chain reaction

Primers D1R/D3Ca

Primer 5: PN/universal

Primer Name: D1R
 Researcher: Alexandra Silva
 Sequence (5' to 3') ACC CGC TGA ATT TAA GCA TA
 Molecular Weight ($\mu\text{g}/\mu\text{mole}$): 6086.0
 Micromolar Extinction Coeff(OD/ μmol) 226.0
 Purity Desalted
 Tm (1 M Na⁺) 64
 Tm (50 mM Na⁺) 43
 % GC 40
 Notes:

Primer Number: M7303B05 (B05)

Primer Length: 20

Scale of Synthesis: 50n mol

μg per OD: 26.9

nmoles per OD: 4.4

OD's 11.10

$\mu\text{g's}^*$ 298.91

nmoles 49.1 491

Coupling Eff. 99%

Primer 7: PN/Universal

Primer Name: D3Ca
 Researcher: Alexandra Silva
 Sequence (5' to 3') ACG AAC GAT TTG CAC GTC AG
 Molecular Weight ($\mu\text{g}/\mu\text{mole}$): 6127.0
 Micromolar Extinction Coeff(OD/ μmol) 225.0
 Purity Desalted
 Tm (1 M Na⁺) 68
 Tm (50 mM Na⁺) 47
 % GC 50
 Notes:

Primer Number: M7303B07 (B07)

Primer Length: 20

Scale of Synthesis: 50n mol

μg per OD: 27.2

nmoles per OD: 4.4

OD's 14.50

$\mu\text{g's}^*$ 394.85

nmoles 64.4 644

Coupling Eff. 99%

Componentes:

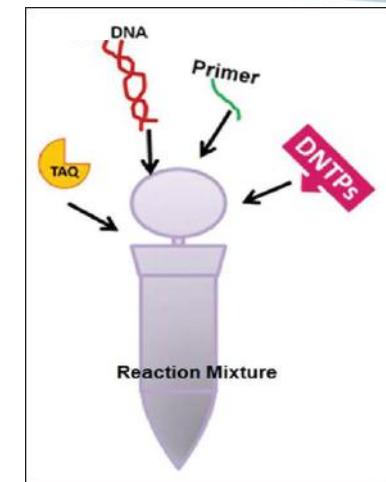
3. Enzima Taq-polimerase

- DNA polimerase
- Origem na bactéria *Thermus aquaticus* que tolera elevadas temperaturas

4. Solução tampão

- MgCl₂ (remove os fosfatos da solução)
Cofactor da Taq
 - Elevadas conc. levam aumento da inespecificidade
 - Baixas concentrações levam diminuição de produto
 - 1.5-2.0 mM

5. Termociclador, equipamento que faz ciclos de temperatura pré-estabelecidos com tempos exatos específicos para cada reação (fragmento a ser amplificado).



PCR-Polymerase Chain reaction

Eppendorf plus



Biorad



TC-plus



Reação de PCR

Ex. GoTaq, Promega

In a sterile, nuclease-free microcentrifuge tube, combine the following on ice:

Component	Final Volume	Final Concentration
5X Green or Colorless GoTaq® Reaction Buffer ¹	10µl	1X (1.5mM MgCl ₂) ²
PCR Nucleotide Mix, 10mM each	1µl	0.2mM each dNTP
upstream primer	Xµl	0.1–1.0µM
downstream primer	Yµl	0.1–1.0µM
GoTaq® DNA Polymerase (5u/µl)	0.25µl	1.25u
template DNA	<u>Zµl</u>	<0.5µg/50µl
Nuclease-Free Water to	50µl	

¹Completely thaw and thoroughly vortex the buffer prior to use.

²More MgCl₂ can be added to the reaction using 25mM MgCl₂ Solution (Cat.# A3511).

PCR-Polymerase Chain reaction

Reação de PCR

Ex. GoTaq, Promega

Table 1. Recommended Thermal Cycling Conditions for GoTaq® DNA

Polymerase-Mediated PCR Amplification. These guidelines are optimal for the Perkin Elmer thermal cycler model 480 or comparable thermal cyclers.

Step	Temperature	Time	Number of Cycles
Initial Denaturation	95°C	2 minutes	1 cycle
Denaturation	95°C	0.5–1 minute	
Annealing	42–65°C*	0.5–1 minute	25–35 cycles
Extension	72°C	1min/kb	
Final Extension	72°C	5 minutes	1 cycle
Soak	4°C	Indefinite	1 cycle

*Annealing temperature should be optimized for each primer set based on the primer T_m .

Reação de PCR GoTaq, Promega

Componentes	Concentrações finais	Nossa concentração stock	Reação base (1 amostra) ul	Nossa reação (1+ Cpos+ Cneg)=3 +1 (erros pipetagem)
Flexi buffer GoTaq	1x	5x	4	
dNTPs	0,2mM	5mM each	1	
MgCl2	1-4mM	25mM	1,5 (1,8mM)	
D1R (Fw)	0,1-1uM	10mM	1 (0.5uM)	
D3Ca (Rv)	0,1-1uM	10mM	1 (0.5uM)	
Taq	1-1.5U	5U/uL	0.1	
DNA template	10pg-1ug	10ng-100ng	1	
H2O	Variável		10.4	
Volume total	-		20ul	

Reação de PCR GoTaq, Promega

Componentes	Concentrações finais	Nossa concentração stock	Reação base (1 amostra) ul	Nossa reação (1+ Cpos+ Cneg)=3 +1 (erros pipetagem)
Flexi buffer GoTaq	1x	5x	4	12
dNTPs	0,2mM	5mM each	1	3
MgCl2	1-4mM	25mM	1,5 (1,8mM)	4,5
D1R (Fw)	0,1-1uM	10mM	1 (0.5uM)	3
D3Ca (Rv)	0,1-1uM	10mM	1 (0.5uM)	3
Taq	1-1.5U	5U/uL	0.1	0.3
DNA template	10pg-1ug	10ng-100ng	1	
H2O	Variável		10.4	31.2
Volume total	-		20ul	

PCR-Polymerase Chain reaction

Programa do termociclador para D1R/D3Ca

Step	Temperature °C	Time	Nº cycles
Desnaturação inicial	95	3 min	1
Desnaturação	95	30 seg	
Emparelhamento	62	35 seg	40
Extensão	72	1 min	
Extensão final	72	10 min	1
Pausa	4	∞	

Reação de PCR

1. Limpar superfícies de trabalho com Lixivia, esperar 10 min e passar por água estéril (lixivia é corrosiva)
Para descontaminar DNA prévio
2. Colocar pipetas de PCR, Eppendorfs, tubos PCR, caneta de acetato na câmara de UV
3. Ligar câmara de UVs 15min
 - Pode realizar-se a mistura fora de uma câmara de PCR
 - UVs destroem o DNA por formar ligações covalentes entre bases
4. Programar o termociclador
5. Gelo num tabuleiro
6. Ordenar todos os componentes que vamos usar (à excepção da taq)
7. Descongelar completamente, mix e spin de todos os componentes (componentes que precipitam no frio)
8. Ir para a Câmara de PCR
9. Marcar tubos de PCR



Reação de PCR

10. Fazer a mix a começar pelos maiores volumes
11. Adicionar a Taq
12. Mix com o dedo
13. Distribuir por todos os tubos 19ul
14. O que restar será no Cneg. Fechar tubo
15. Colocar o DNA nos tubos
16. Dirigir à sala pós-PCR
17. Colocar no termociclador (não retirar as escorras)
18. Run



PCR Master mix

- Buffer + MgCl₂ + Taq
- faltam primers, DNA, H₂O
- Há com e sem Loading Dye



Vantagens MIX

- Menos erros pipetagem
- Processo mais rápido

Desvantagem

- Não se pode optimizar condições de Magnésio ou dNTPs
- Se contaminação, há maior desperdício

Direct PCR Master mix

Phire tissue Master Mix – “Extração de DNA” + PCR directo

6.1 Solid samples

Animal tissues

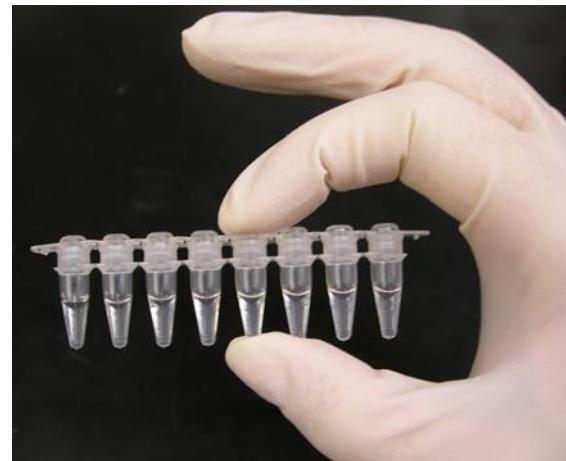
1. **Direct protocol:** Take a sample of 0.35-0.5 mm in diameter from tissue with a sterile scalpel (or small peace, e.g. one *Drosophila* leg) or by a tissue puncher. Place the sample directly into the PCR reaction (50 µL of volume). It is recommended to place the sample into the liquid rather than into an empty tube. Make sure that you see the sample in the solution.
2. **Dilution & storage protocol:** Before beginning, warm a heat block to 98 °C. Place the tissue sample into 20 µL of Dilution Buffer. Add 0.5 µL of DNARelease Additive. Mix by vortexing the tube briefly, and spin down the solution. If a larger sample is used, adjust the volume of the Dilution Buffer and DNARelease Additive accordingly. Make sure the sample is covered with the solution. Incubate the reaction for 2–5 minutes at room temperature and then place the tube into the pre-heated (98 °C) block for 2 minutes. Spin down the remaining tissue and store the supernatant at –20 °C if not used immediately. Usually 1 µL of supernatant is sufficient for a 20 µL PCR reaction. In some cases the supernatant may have to be diluted 1:10 or 1:100, or the PCR reaction performed in a 50 µL volume.

Direct PCR Master mix

Phire tissue Master Mix – “Extração de DNA” + PCR directo

Table 1. Pipetting instructions

Component	20 µL rxn	50 µL rxn*	Final conc.
H ₂ O	add to 20 µL	add to 50 µL	-
2X Phire Tissue Direct PCR Master Mix	10 µL	25 µL	1X
Primer A	X µL	X µL	0.5 µM
Primer B	X µL	X µL	0.5 µM
Sample (see Section 6) Direct protocol:	-	Amount depends on the sample**	-
Dilution & Storage protocol:	0.5-1 µL	2.5 µL	



*50 µL reaction volume is recommended for the direct protocol.

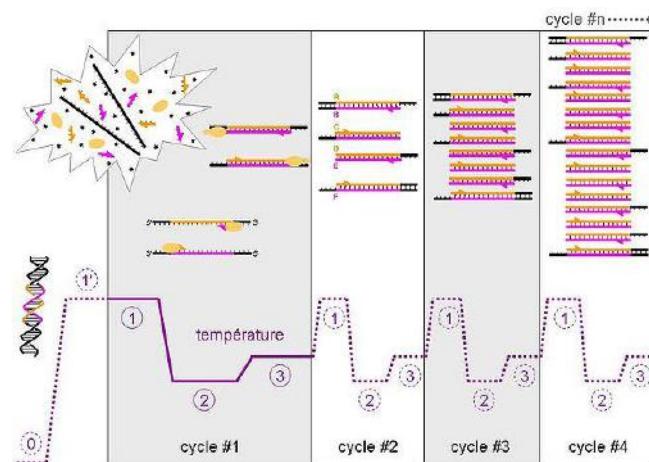
**0.5 mm punch or a small sample of tissue
(see www.thermoscientific.com/directpcr)

Programa de PCR

1. Desnaturação inicial

Fundamental, pois compromete todos os passos seguintes

1-3 min a 95ºC - GC content é 50% ou menos
até 10min - GC-rich templates



Programa de PCR

2. Melting/Desnaturação

quebra das pontes de hidrogénio
separação da dupla cadeia de DNA
0.5-2 min entre 94 a 96°C

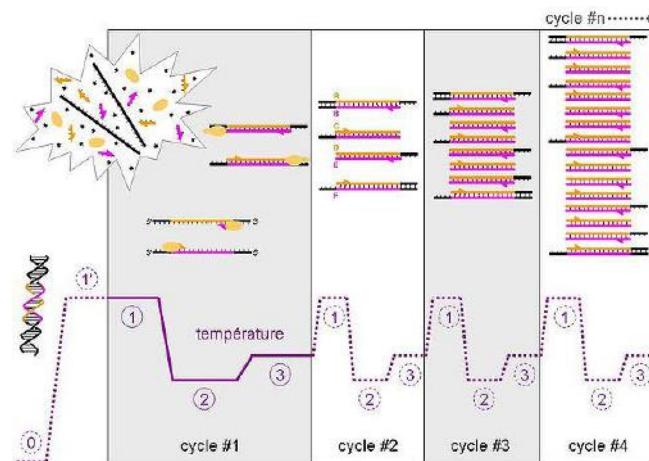
3. Annealing/Emparelhamento

Primers emparelham com cadeia molde
Depende da quantidade de citosina (C) e
guanina (G)

5°C abaixo temperatura melting primer-template
0.5-2min entre 50 a 60 °C
Se produtos inespecíficos-aumentar com
incrementos de 1-2°C

4. Extension/Extensão

A enzima sintetiza a nova molécula
70-75°C
1min até 2Kb
Quando fragmentos maiores – adiciona
1min/1000bp



Programa de PCR

5. Ciclos

Depende da quantidade de DNA template e do nº de cópias que pretendo no final

25 a 40 ciclos

taxa de replicação é exponencial

6. Extensão final

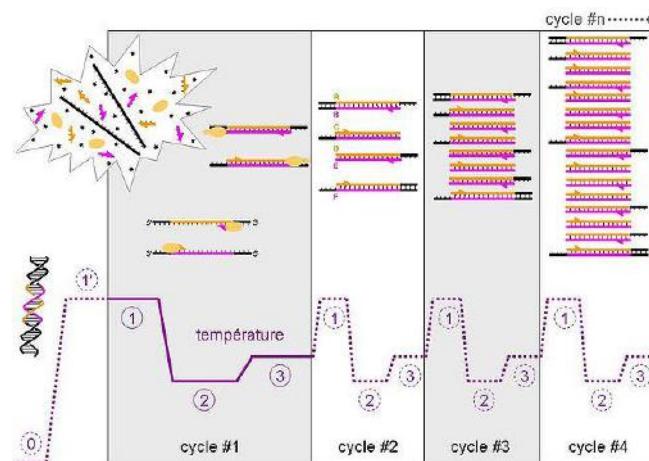
72°C – 5-15min

terminar o final das cadeias

Taq adiciona poly-A tails ao extremo 3' (importante para a clonagem)

7. Final hold

4°C - ∞



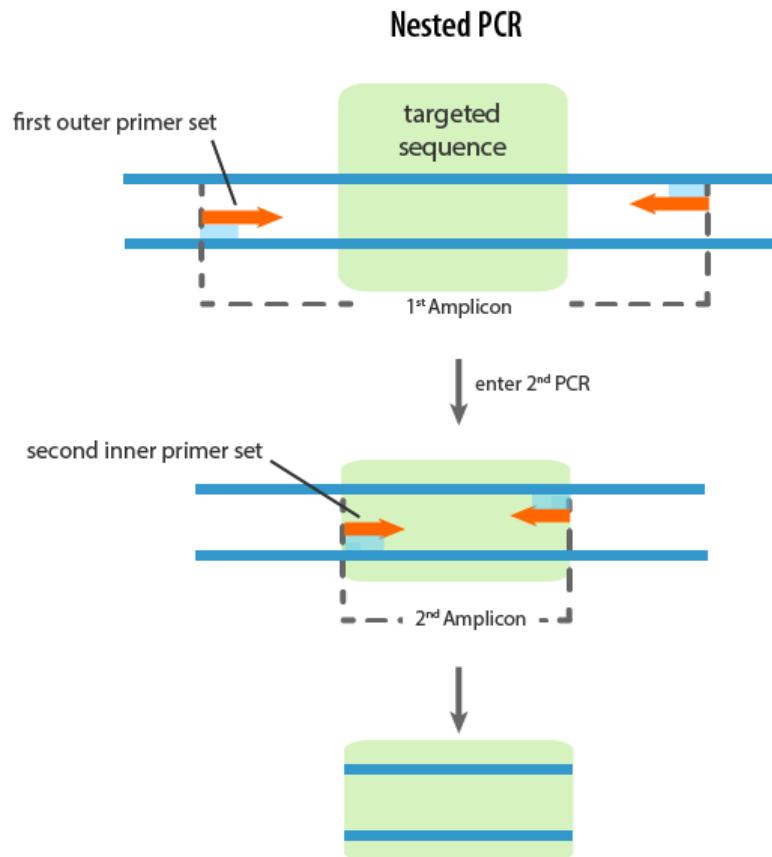
PCR-Polymerase Chain reaction

PCR additives

- **DMSO:** Reduce **secondary structure** that could inhibit the progress of the polymerase. Especially useful for GC rich templates. Use at a final concentration of 5-10%.
- **Glycerol:** Reduces **secondary structure**. Use 5-10%
- **Betaine monohydrate:** Reduces **secondary structure**. Use 1 to 3M
- **BSA:** Very useful for templates that may be contaminated with **humic acids** (e.g. environmental samples contaminated with soil) and is also reported to prevent reaction components from sticking to the tube wall. Use up to 0.8 mg/ml
- **Tween-20:** Can neutralize **SDS** left over from template DNA preparation that would inhibit the reaction. Use 0.25 to 1% final concentration
- **Formamide:** Increases the **stringency** of primer annealing, resulting in less non-specific priming and increased amplification efficiency. Use 1-10%
- **Tetramethyl ammonium chloride:** Similar to formamide. Use 10-100mM
- **7-deaza-2'-deoxyguanosine:** A dGTP analogue that is especially useful for extremely **GC rich templates**. Success is reported with up to 83% GC. Use a 1:3 ratio of dGTP:7-deaza-2'-deoxyguanosine

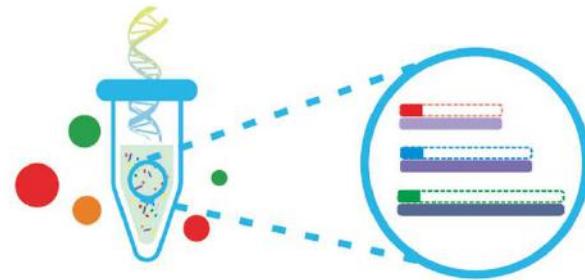
Nested PCR

- Aumentar a especificidade do PCR
- 2 pares de primers para o mesmo locus
 - 1 par mais externo (1^a amplificação)
 - Outro par mais interno (restantes amplificações)
- Adicionar-se em 2 reações de PCR consecutivas



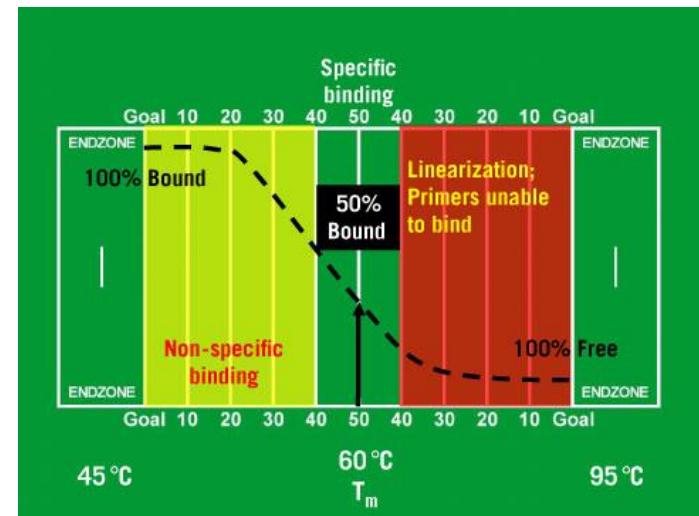
Multiplex PCR

- Vários alvos numa única reação PCR
- Mesma mistura de PCR
 - 1 template
 - Vários pares de primers
- Normalmente os produtos amplificados têm vários tamanhos
- Temperaturas de annealing e desenho de primers têm de ser optimizados para todos trabalharem correctamente numa mesma reação



Touchdown PCR

- Reduz amplificações não-específicas, pelo decréscimo da temperatura de annealing
- Início- 3-5°C acima T_m
 - Elevada especificidade
- PCR continua – temp. em cada ciclo decresce 0.2°C até chegar 3-5°C abaixo da T_m .
- Temperaturas elevadas – elevada especificidade
- Temperaturas baixas – amplificações mais eficientes, do produto formando nos 1ºs ciclos
- Logo a 1ª sequência amplificada é a única entre as regiões de maior especificidade do primer e será a mais abundante no final do PCR



Hot-start PCR

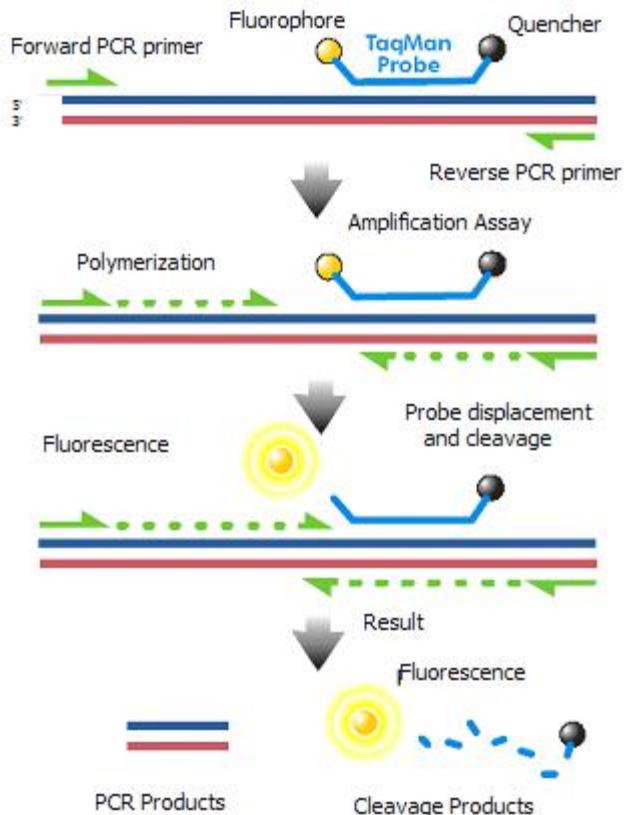
- A reação de PCR é ativada quando a temperatura atinge 94°C
- Aumenta a especificidade do PCR, pois a DNA polimerase contém um anticorpo, que se desnatura e ativa a enzima ao atingir a temperatura de 94°C.
- DNA polimerases que não possuem este inibidor podem amplificar produtos inespecíficos à temperatura ambiente.

<https://youtu.be/ID6KY1QBR5s>

<https://youtu.be/mvvP90Cpdfc>

Real-time PCR (qPCR)

- PCR que quantifica o DNA obtido por fluorescência
- Resultados mais rápidos e mais precisos
- Detecção SyberGreen
 - Corante não específico que se intercala na molécula de DNA de cadeia dupla
- Detecção sondas Taqman (ex.)
 - Sequencias específicas complementares ao meu alvo, tipo “primers” marcadas com uma molécula fluorescente que permite deteção depois da hibridação



PCR convencional	Real-time PCR
Menor precisão	Maior
Menor sensibilidade	Maior
Menor resolução	Maior
Não automatizado	Automatizado
Resultados não numéricos	Numéricos
Possibilidade de produtos inespecíficos	Não influenciado por amplificação não específica
Amplificação monitorizada no final	Amplificação monitorizada em tempo real
Gel de agarose	Não há processamentos pós-pcr
Menor rapidez na corrida	Maior
Maior quantidade DNA	Requer 1000X menos DNA
Coleta de dados na fase final	Coleta de dados na fase exponencial



Preparação de amostras

Resultados

2016Mat03

Primers:

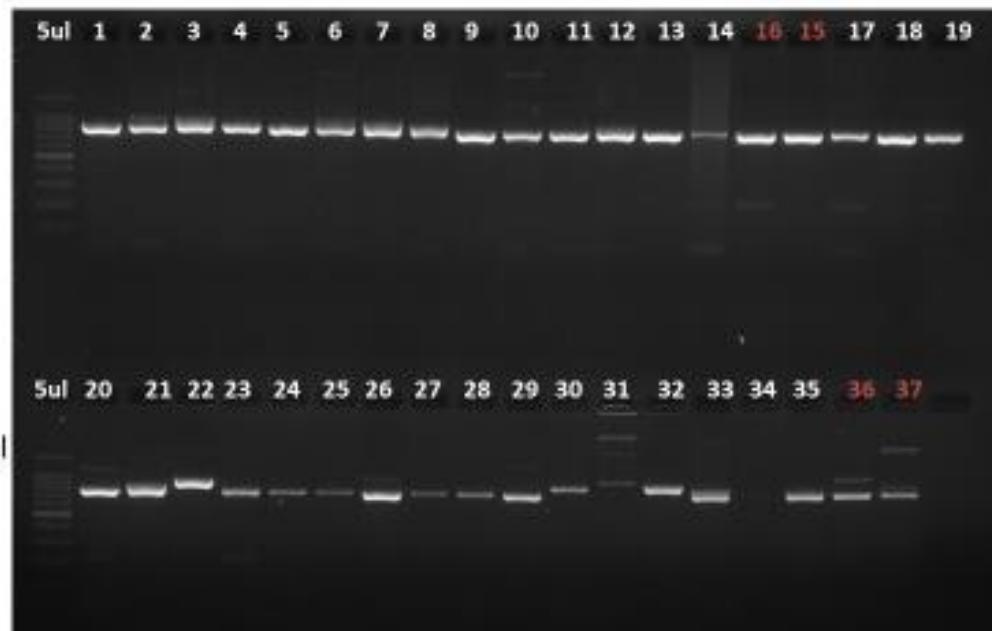
D1R / D3Ca

ver amostras caderno ou slide seguinte

95°C-3min
94°C-30''
62°C-35'' - 40x
72°C-1min
72°C-10min
4°C-∞

H2O-7
Phire-10
D1r-1
D3Ca-1
DNA-1

2% Agarose- 100ml
de gel
10W-40min
5 ul pcr
4 greensafe



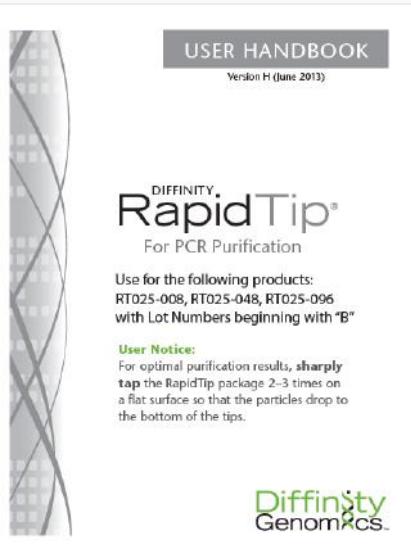
Preparação de amostras

Purificação de produtos de PCR

Eliminam primers, dNTPs, enzimas, produtos PCR pequenos e inespecíficos e sais de fragmentos de PCR >100 bp, que podem afectar as aplicações seguintes

- Silica based columns (mais comum)
 - Sal caotrópico (quebra pontes de H) desnatura o DNA que se liga à resina de sílica da coluna, permitindo que se separe dos restantes componentes da amostra
 - Depois da lavagem, o DNA é eluído com uma solução baixa em sal, que permite que se reorganize e perca afinidade com a membrane
 - Eluir em H₂O (não afecta procedimentos subsequentes, menos estável). Eluir em Tris-EDTA (mais estável, mas sais podem afectar procedimentos subsequentes)

Purificação de produtos de PCR



Pipetting Examples



A. Correct mixing:
Sample and particles mixing well.



B. Incorrect mixing:
Particles separated by air bubble.

RapidTip PCR Purification Protocol

Prepare Samples:

Diffinity RapidTip is optimized for a 25 µl PCR reaction size.

RT025-008, 048, 096
Purifies 20-30 µl sample size.
For PCR volumes >30 µl, aliquot 25 µl into a tube.
For PCR volumes <20 µl, please dilute to 25 µl.

Dilution works best with highly concentrated DNA samples (>50 ng/µl) as your sample concentration will be reduced.

Prepare Tips:

Diffinity RapidTip for PCR Purification contains proprietary particles that purify a PCR reaction; these particles can adhere to the pipette tip walls during shipping. For optimal results, sharply tap the box 2-3 times on a flat surface so that particles are at the bottom of the tips (near the retainer).

Please note that it is normal to see fine dust-like particles on the side of the tip. After tapping the box, you should expect to see about 1-2mm of white particles above the retainer at the small end of the pipette tip.

Purify Samples:

1. For $25 \pm 5 \mu\text{l}$ samples, program pipettor to aspirate 30 µl.
2. Place Diffinity RapidTip on pipettor — single or multichannel.

Please note that you can use a multi-channel pipettor to mix more than 1 sample at a time for even higher productivity.

Preparação de amostras

Purificação de produtos de PCR

RapidTip PCR Purification Protocol

Prepare Samples:

Diffinity RapidTip is optimized for a 25 µl PCR reaction size.

RT025-008, 048, 096
Purifies 20-30 µl sample size.
For PCR volumes >30 µl, aliquot 25 µl into a tube.
For PCR volumes <20 µl, please dilute to 25 µl.

Dilution works best with highly concentrated DNA samples (>50 ng/µl) as your sample concentration will be reduced.

Prepare Tips:

Diffinity RapidTip for PCR Purification contains proprietary particles that purify a PCR reaction; these particles can adhere to the pipette tip walls during shipping. For optimal results, sharply tap the box 2–3 times on a flat surface so that particles are at the bottom of the tips (near the retainer).

Please note that it is normal to see fine dust-like particles on the side of the tip. After tapping the box, you should expect to see about 1–2mm of white particles above the retainer at the small end of the pipette tip.

Purify Samples:

1. For 25 ± 5 µl samples, program pipettor to aspirate 30 µl.
 2. Place Diffinity RapidTip on pipettor — single or multichannel.
- Please note that you can use a multi-channel pipettor to mix more than 1 sample at a time for even higher productivity.*

3. Place pipette tip into PCR sample solution.

4. Pre-wet the particles on first aspirate (see page 10).

Aspirate about half the sample and then pause for 5 seconds to ensure complete wetting of the particles before mixing. Aspirate the rest of the sample and then dispense.

5. Set timer and mix for 60 seconds (approximately 15 aspirate/dispense cycles).

Please note that pipetting will be slower than normal — wait for liquid to completely fill the tip to begin the next mix. It is not necessary to drive liquid completely out of the tip on every dispense. The particles should mix completely with the solution and make it appear cloudy while inside the tip.

6. On the final dispense, you can use your pipettor's blowout mode for maximum liquid recovery.

Your purified PCR Amplicon is now ready for downstream Sanger amplification.

Preparação de amostras

Purificação de produtos de PCR

5.2 Cutting Gels

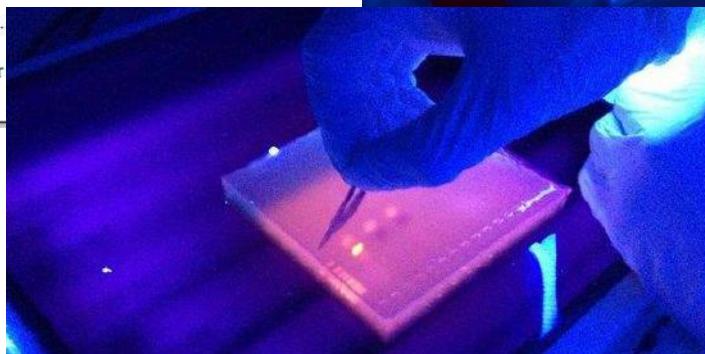
1. Open the drawer until it is fully extended and the Trans UV LED stops blinking.
2. Place the clear acrylic UV protection screen in the horizontal slot in the inside of the drawer facing the transilluminator.



!! CAUTION!! !! VORSICHT !!

WARNING: EXPOSURE TO UV RADIATION IS HAZARDOUS TO HEALTH. PLEASE WEAR PROPER UV PROTECTIVE CLOTHING AND FACE AND EYE SHIELDS WHEN CUTTING GELS. THE UV SHIELD PROVIDED WITH THE SYSTEM IS NOT ADEQUATE PROTECTION AGAINST UV GENERATED BY THIS SYSTEM.

3. Press the **Trans UV** button to turn on the UV illumination.
4. Cut the gel.
5. Press the **Trans UV** button and the UV transilluminator protection shield before closing the drawer.



Preparação de amostras

Purificação de produtos de PCR



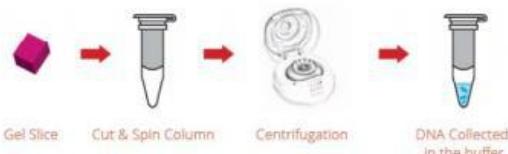
Cut&Spin Gel Extraction Columns

#GS131.0250: 250 columns, #GS131.0500: 500 columns

Cut&Spin Gel Extraction Columns are ready-to-use spin columns for the extraction of DNA from agarose gels.

Protocol

Just cut out the desired DNA-band from the agarose gel, place it on the top of the column media, and centrifuge at 5.000-6.000g at room temperature for only 5-10 minutes.



Supplied Material

Spin columns are supplied in bags of 10, inserted in a 1,5ml receiver tube and closed with an attached receiver tube lid.

Preparação de amostras

Send to sequence

- Macrogen
- Stabvida
- GATC (Nzytech)
- ...

Mandar sequenciar à **GATC**
(intermediário Nzytech: diogo.comprido@nzytech.com)

LightRun (LR)

1. → Preparação das amostras
 - a. → Purificar.
 - b. → Tubo 1 — 10ul exactamente (tubo de 1,5ml)
5ul DNA (20-80ng/ul) + 5ul Primer (5uM)
ou 9ul DNA + 1ul Primer (desde que a conc. DNA seja entre 100 a 400ng e o primer 25pmol)
 - c. → Se quiser sequenciar o reverse tenho de criar mais um tubo com PCR product e primer reverse
2. → Labelling das amostras
 - a. → Colar um barcode branco ao tubo (Light Run for Tubes)
 - b. → Se quiser o Reverse, colo outro barcode
 - c. → Colar o outro autocolante no caderno para meu registo
3. → Preparar envelope (?)
 - a. → Imprimir ou guardar assim que clica no label
 - b. → Colar num envelope/saco
4. → Chamar UPS — 707 232323
 - i. → Recolha etiqueta pré-paga — UPS save!
 - ii. → Até 1kg
 - iii. → Alemanha
 - iv. → Morada
IPMA - Fitoplâncton Lab.
Rua Alfredo Magalhães Ramalho, 6
1495-006 Lisboa

Preparação de amostras

Send to sequence

Macrogen
Stabvida
GATC (Nzytech)

Mandar sequenciar à GATC
(intermediário-Nzytech-diogo.comprido@nzytech.com)

SupremeRun (SR)

- 1.→ Preparação das amostras
 - a.→ Não preciso de purificar. Está incluído no preço
 - b.→ DNA sample-20ul exactamente, com 10-50ng/ul, no tubo 1 (tubo de 1,5ml)
 - c.→ Primer FW-20ul exactamente, com 10uM (o que usamos), no tubo 2
 - d.→ Primer Ry-20ul exactamente, com 10uM (o que usamos), no tubo 3
 - i.→ Estes 20ul dão para 6 rxns
- 2.→ Labelling das amostras
 - a.→ Colar um barcode verde ao tubo 1 (GATC-DNA)
 - b.→ Colar outro barcode amarelo ao tubo 2 (GATC-Primer)
 - c.→ Colar outro barcode amarelo ao tubo 3 (GATC-Primer)
- 3.→ Ir à myGATC account
 - a.→ Sanger sequencing
 - b.→ Supreme Run order
 - c.→ Tubes
 - d.→ PCR fragments
 - e.→ Can be purified by GATC
 - f.→ Select DNA Barcode
 - g.→ Nomear a minha amostra
 - h.→ Select Primer Barcode
 - i.→ Não preciso de colocar sequência
 - j.→ Clicar + Reuse DNA
- 4.→ Preparar envelope
 - a.→ Imprimir ou guardar assim que clica no label
 - b.→ Colar num envelope/saco
- 5.→ Chamar UPS - 707-232323
 - i.→ Recolha etiqueta pré-paga - UPS saver
 - ii.→ Até 1kg
 - iii.→ Alemanha
 - iv.→ Morada
 - IPMA- Fitoplâncton Lab.
 - Rua Alfredo Magalhães Ramalho, 6
 - 1495-006 Lisboa

Preparação de amostras

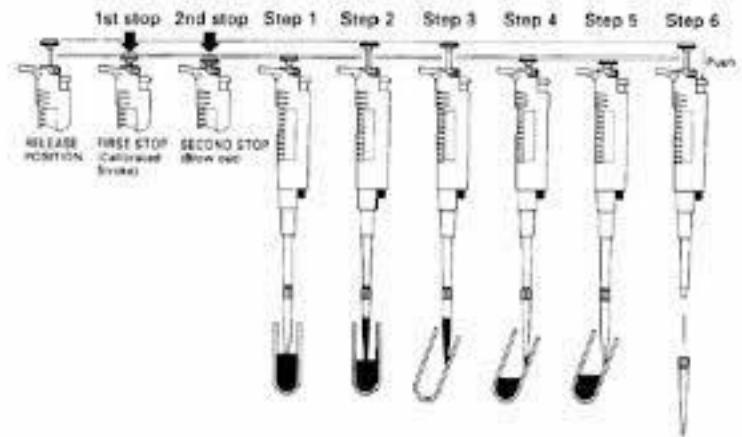
Micropipetas

Como pipetar

1. Aspirar lentamente e na vertical
2. Descartar lentamente e com ângulo de 40º, com a ponta encostada ao recipiente, até à 1ª posição
3. Carregar até à 2ª posição, arrastando a ponta pelo recipiente

Manutenção e Limpeza das Micropipetas

- Manter a pipeta sempre em posição vertical e no volume máximo (a pipeta deitada de um dia para o outro é suficiente para a descalibrar. A mola em tensão pode provocar também uma descalibração)
- Para encaixar as pontas, não bater. Rodar nas monocanais e encaixar frente e trás nas pluricanais.
- Limpar externamente a mp com alcool a 70º, sempre que necessário.
- Limpeza interna - 5 cargas/descargas **sem** ponta, com alcool a 70º, e deixar a secar no fim de semana ou 30min na estufa até 60ºC, se houver urgência
- Lixívia e autoclavagem para situações mais complicadas





INTRODUÇÃO À BIOLOGIA MOLECULAR E BIOINFORMÁTICA: Bioinformatics

Lisboa, 1-3 Junho

João Machado

Bárbara Frazão

Bioinformatic

- What is it? Bioinformatics is the creation , development and operation of databases and other computational tools to collect, organize and interpret data
- Data Sources ? They are usually derived from biological data experiences that provide quantitative and qualitative data



Use of databases in bioinformatics as repositories and sources of information

Data Warehouses

- From 1982 databases began to be created for storing information and sequences of nucleotides

Examples

- European Molecular Biology Laboratory:
<http://www.embl.org/> (Europe)
- National Institutes of Health:
<http://www.ncbi.nlm.nih.gov> (North America)
- DNA Databank (DDBJ):
<http://www.ddbj.nig.ac.jp/> (Japan)

Data Warehouses

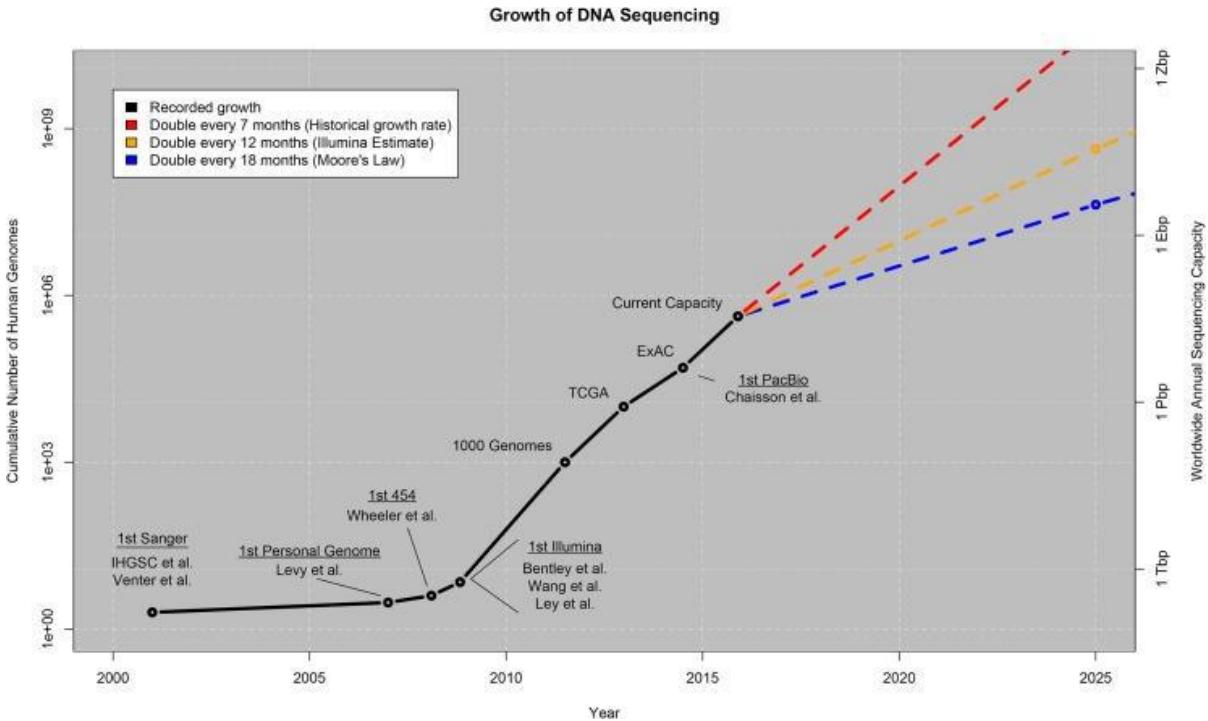
- From 1986 databases of amino acids (proteins)

Exemplos

- Swissprot/TrEMBL
- PIR
- In 2002 joined in UniProtKb
(<http://www.uniprot.org/>)
- **UniProtKB/Swiss-Prot** which is manually annotated and is reviewed and
- **UniProtKB/TrEMBL** which is automatically annotated and is not reviewed.

Data growth

- “100-gigabase” in August 2005. 200 billion bp in September 2007. The amount of data doubles every 18 months.



PLoS Biol. 2015 Jul; 13(7): e1002195.

Conventions

GenBank gb|accession.version

EMBL emb|accession.version

DDBJ dbj|accession.version

NCBI RefSeq ref|accession.version

PDB pdb|entry|chain

Patents pat|country|number

NBRF PIR pir||entry

SWISS-PROT sp|accession|entry

Protein Research Foundation prf|name

GenInfo Backbone Id bbs|number

General database identifier gnl|database|identifier

Local Sequence identifier lcl|identifier

Conventions

RefSeq categories

Experimentally determined and curated		Genome annotation (computational predictions from DNA)	
NC	Complete genomic molecules		
NG	Incomplete genomic region		
NM	mRNA	XM	Model mRNA
NR	RNA (non-coding)		
NP	Protein	XP	Model protein

Conventions

IUB/IUPAC nucleotide and ambiguity codes

A	adenosine	M	A or C (amino)	V	A, C, or G
C	cytidine	K	G or T (keto)	H	A, C, or T
G	guanine	R	A or G (purine)	D	A, G, or T
T	thymidine	Y	C or T (pyrimidine)	B	C, G, or T
U	uridine	S	A or T (strong)	–	Gap of indeterminate length
		W	C or G (weak)	N	A, C, G, or T (any or unknown)

R	A,G
Y	C,T
M	A,C
K	G,T
S	C,G
W	A,T
H	A,C,T
B	C,G,T
V	A,C,G
D	A,G,T
N	A,C,G,T

Conventions

- Gene Names (<http://www.genenames.org/>)

HGNC
HUGO Gene Nomenclature Committee

Search everything ▾ Search symbols, keywords or IDs

Use * to search with a root symbol (eg ZNF*)

Home Downloads Gene Families Tools Useful links About Newsletters Contact Us Help Request symbol

FAQ

[What is the HGNC?](#)
[What is HGNC-approved nomenclature and why do we need it?](#)
[Where can I find information about existing human gene symbols?](#)
[What is a stem symbol?](#)
[Where can I find the Nomenclature Guidelines?](#)
[Do I have to use the approved symbols?](#)
[How should I cite HGNC nomenclature resources?](#)

Latest News

[Proposed change to the custom download tool \(give us your feedback\)](#) 

We are proposing simplifying our "Custom Downloads" tool by bringing the data provided in line with that displayed in our symbol reports. Currently users can download two separate fields for some IDs: "HGNC curated" and "mapped data". This has caused some confusion as in our symbol reports HGNC curated data are displayed in preference, and mapped data are only shown if there is no HGNC curated ID i.e. only one ID is shown per symbol. Please use our [feedback](#) form to comment and let us know if this update may

File Formats

- Fasta files

```
>seq1
-----KSKERYKDENGNYFQLREDWW DANRE
>seq2
-----YEGLTTANGXKEYYQDKNGGNFFKLREDWW TANRE
>seq3
-----SQRHYKD-DGGNYFQLREDWW TANRH
>seq4
-----NVAALKTRYEK-DGQNFYQLREDWW TANYF
```

File Formats

- **Phylip interleaved**
- The first line of the input file contains the number of species and the number of characters separated by blanks. The information for each species follows, starting with a ten-character species name (which can include punctuation marks and blanks), and continuing with the characters for that species. Phylip format files can be interleaved, as in the example below, or sequential.

```
4 123
seq1 ----- ---KSKERYK DENGNYFQL REDWWDANRE
seq2 ----- YEGLT TANGXKEYYQ DKNGGNFFKL REDWWTANRE
seq3 ----- SQRHYK D-DGGNYFQL REDWWTANRH
seq4 ----- NVAALKTRYE K-DGQNFYQL REDWWTANRA

TVWKAITCNA --GGGKYFRN TCDG--GQNP TETQNNCRCIG-----
TVWKAITCGA P-GDASYFHA TCDSGDGRGG AQAPHKCRCG G-----
TVWEAITCSA DKGNA-YFRR TCNSADGKSQ SQARNQCRC- --KDENGKN-
TIWEAITCSA DKGNA-YFRA TCNSADGKSQ SQARNQCRC- --KDENGXN-
```

File Formats

- Phylip sequencial

4 123

```
seq1 ----- KSKERYK DENGNYFQL REDWWDANRE
TVWKAITCNA --GGKYFRN TCDG--GQNP TETQNNCRCI G-----
seq2 ----- YEGLT TANGXKEYYQ DKNGGNFFKL REDWWTANRE
TVWKAITCGA P-GDASYFHA TCDSGDGRGG AQAPHKCRCG G-----
seq3 ----- SQRHYK D-DGGNYFQL REDWWTANRH
TVWEAITCSA DKGNA-YFRR TCNSADGKSQ SQARNQCRC- --KDENGKN-
seq4 ----- NVAALKTRYE K-DGQNFYQL REDWWTANRA
TIWEAITCSA DKGNA-YFRA TCNSADGKSQ SQARNQCRC- --KDENGXN-
```

File Formats

- Nexus

```

#NEXUS
BEGIN DATA;
    DIMENSIONS NTAX=10 NCHAR=22;
    FORMAT MISSING=? DATATYPE=DNA GAP=- EQUATE="0=A 1=C";
    OPTIONS GAPMODE=MISSING;
MATRIX

[00000000000000000000]
[00000000000000000000]
[00000000011111111222]
[1234567890123456789012]

TaxonA          AAAAAAAAAAAAAAAA000000
TaxonB          AA----AAA--AAA1--100
TaxonC          AAA-AAAAAA--AAA010010
TaxonD          AAAGAA-AAAAGAA-A001001
TaxonE          ACGTACGTACGTACGT000000
TaxonF          AAAAAAAA000000
TaxonG          AA----AAA--AAA1--100
TaxonH          AAA-AAAAAA--AAA010010
TaxonI          AAAGAA-AAAAGAA-A001001
TaxonJ          ACGTACGTACGTACGT000000
;

END;

[ Indel Character      Sequence Region ]
[ -----           ----- ]
[                               ]
[ 17                 3-7           ]
[ 18                 4-5           ]
[ 19                 7-7           ]
[ 20                 11-12          ]
[ 21                 12-13          ]
[ 22                 15-15          ]

```

File Formats

GenBank

			FEATURES	Location/Qualifiers
LOCUS	AF023787	618 bp	DNA linear	PLN 02-MAY-1998
DEFINITION	Bryum stenotrichum small ribosomal protein 4 (rps4) gene, chloroplast gene encoding chloroplast protein, partial cds.		source	1..618 /organism="Bryum stenotrichum" /organelle="plastid:chloroplast"
ACCESSION	AF023787			/mol_type="genomic DNA"
VERSION	AF023787.1	GI:3098167		/db_xref="taxon:66994"
KEYWORDS	.		gene	<1..573 /gene="rps4"
SOURCE	chloroplast Bryum stenotrichum			<1..573 /gene="rps4"
ORGANISM	<u>Bryum stenotrichum</u>		CDS	<1..573 /gene="rps4" /codon_start=1 /product="small ribosomal protein 4" /protein_id="AAC15532.1" /db_xref="GI:3098168"
Eukaryota; Viridiplantae; Streptophyta; Embryophyta; Bryophyta; Moss Superclass V; Bryopsida; Bryidae; Bryanae; Bryales; Bryaceae; Bryum.				/translation="RRLGSLPGLTNKTPQLKTNNSINQSISNKKISQYRIRLEEKQKLR FHYGITERQLNYVRIARKAKGSTGEVLLQLEMRLDNVIFRLGMAPTIPGARQLVNH RHILVNDRIVNIPSYRCKPEDSITIKDRQKSQAIISKNLNLQKYKTPNHLTYNFNLKK KGLVNQILDRESIGLKINELLVVEYYSRQA"
REFERENCE	1 (bases 1 to 618)			
AUTHORS	Cox,C.J. and Hedderson,T.A.J.			
TITLE	Phylogenetic relationships among the ciliate arthrodontous mosses: evidence from chloroplast and nuclear DNA sequences			
JOURNAL	Unpublished			
REFERENCE	2 (bases 1 to 618)			
AUTHORS	Cox,C.J. and Hedderson,T.A.J.			
TITLE	Direct Submission			
JOURNAL	Submitted (11-SEP-1997) Dept. of Botany, School of Plant Sciences, University of Reading, Whiteknights, Reading, Berkshire RG6 6AS, United Kingdom			
			ORIGIN	
				1 cgcgcgttag gatctttacc aggactaact aataaaacac cccagttaaa aactaattcg 61 atcaatcaat caatatctaa taaaaaaaatt tctcaatatc gcattcgtt ggaagaaaaa 121 caaaaattac gtttccatta tggataaca gagcgacaat tacttaatta tgtacgtatt 181 gctagaaaaag ctaaagggtc aacaggtgaa gtcttattac aattactga aatgcgccta 241 gataacgta ttttcgatt agtgtatggct cctacaattc ctggagcaag gcaactagta 301 aatcatagac atattttagt taatgatcgt atagtaataa taccagaat tccgtgtaaa 361 cctgaggatt ctattactat aaaagatcga caaaaatctc aggctataat tagaaaaat 421 ttaaatttgt ataaaaata taaaacacca aatcattaa cttataattt tttaaaaaaa 481 aaaggatgg ttaatcaaact actagatcgt gaatccattt gttttaaaat aaatgattha 541 ttatgttag aatattatttc tcgccaagct taattaacaa ctaagagtgat ttgtatatt 601 atacataata aaaaatttg

File Conversions

- Software (e.g. Seaview, Mega, Mesquite, Bioedit, etc)
 - Requires instalation
- Scripts (generally in Python or Perl)
 - Several are freely distributed in github
(<https://github.com/>)
- Web-based tool (easy)

Hands-On 1

- <https://goo.gl/VFwldq>
- 1) Descarregar as sequencias na pasta hands_on_1
 - 2) Usar o site: http://www.ebi.ac.uk/Tools/sfc/emboss_seqret/
 - 3) Analisar o formato original das sequencias
 - 4) Converter as sequencias para Fasta
 - 5) Adicionar num ficheiro separado usando editor de texto



Search Methods

Search NCBI

[NCBI](#) [Resources](#) [How To](#)

[Sign in to NCBI](#)

NCBI
National Center for
Biotechnology Information

All Databases ▾ [Search](#)

[NCBI Home](#)

[Resource List \(A-Z\)](#)

[All Resources](#)

[Chemicals & Bioassays](#)

[Data & Software](#)

[DNA & RNA](#)

[Domains & Structures](#)

[Genes & Expression](#)

[Genetics & Medicine](#)

[Genomes & Maps](#)

[Homology](#)

[Literature](#)

[Proteins](#)

[Sequence Analysis](#)

[Taxonomy](#)

[Training & Tutorials](#)

[Variation](#)

Welcome to NCBI

The National Center for Biotechnology Information advances science and health by providing access to biomedical and genomic information.

[About the NCBI](#) | [Mission](#) | [Organization](#) | [NCBI News](#) | [Blog](#)

Submit

Deposit data or manuscripts into NCBI databases



Download

Transfer NCBI data to your computer



Learn

Find help documents, attend a class or watch a tutorial



Develop

Use NCBI APIs and code libraries to build applications



Analyze

Identify an NCBI tool for your data analysis task



Research

Explore NCBI research and collaborative projects



Popular Resources

[PubMed](#)

[Bookshelf](#)

[PubMed Central](#)

[PubMed Health](#)

[BLAST](#)

[Nucleotide](#)

[Genome](#)

[SNP](#)

[Gene](#)

[Protein](#)

[PubChem](#)

NCBI Announcements

NCBI launches new Twitter account for NCBI Bookshelf 23 May 2016

NCBI has a new Twitter feed - [@nchbooks](#) - to announce new books

New NCBI Insights blog post: Fast Sequence Inspection with ORFfinder and SmartBLAST (PubMed Labs) 16 May 2016

The latest book on NCBI Insights

RefSeq release 76 is now available 18 May 2016

RefSeq release 76 is accessible online, via FTP and through NCBI's programming utilities. This full release

[More...](#)

Search NCBI

NCBI Resources How To Sign in to NCBI

Search NCBI databases

Help

Results found in 32 databases for "pp2a"

Literature			Genes		
Books	102	books and reports	EST	1,681	expressed sequence tag sequences
MeSH	27	ontology used for PubMed indexing	Gene	18,715	collected information about gene loci
NLM Catalog	3	books, journals and more in the NLM Collections	GEO DataSets	32	functional genomics studies
PubMed	3,887	scientific & medical abstracts/citations	GEO Profiles	12,292	gene expression and molecular abundance profiles
PubMed Central	11,565	full-text journal articles	HomoloGene	87	homologous gene sets for selected organisms
Health			PopSet	35	sequence sets from phylogenetic and population studies
ClinVar	6	human variations of clinical significance	UniGene	263	clusters of expressed transcripts
dbGaP	0	genotype/phenotype interaction studies	Proteins		
GTR	0	genetic testing registry	Conserved Domains	32	conserved protein domains
MedGen	7	medical genetics literature and links	Protein	110,089	protein sequences
OMIM	81	online mendelian inheritance in man	Protein Clusters	34	sequence similarity-based protein clusters
PubMed Health	0	clinical effectiveness, disease and drug reports	Structure	194	experimentally-determined biomolecular structures
Genomes			Chemicals		
Assembly	0	genome assembly information	BioSystems	6,046	molecular pathways with links to genes, proteins and chemicals
BioProject	25	biological projects providing data to NCBI	PubChem BioAssay	5,395	bioactivity screening studies
BioSample	0	descriptions of biological source materials	PubChem Compound	3	chemical information with structures, information and links
Clone	5	genomic and cDNA clones	PubChem Substance	178	deposited substance and chemical information
dbVar	995	genome structural variation studies			
Epigenomics	0	epigenomic studies and display tools			
Genome	2	genome sequencing projects by organism			
GSS	24	genome survey sequences			
Nucleotide	153,499	DNA and RNA sequences			
Probe	215	sequence-based probes and primers			
SNP	7,538	short genetic variations			
SRA	5	high-throughput DNA and RNA sequence read archive			
Taxonomy	0	taxonomic classification and nomenclature catalog			

Search NCBI

NCBI Resources How To Sign in to NCBI

Gene Gene pp2a Create alert Advanced Search

Gene sources Gene sources

Genomic Genomic

Organelles Organelles

Plasmids Plasmids

Categories Categories

Alternatively spliced Alternatively spliced

Annotated genes Annotated genes

Non-coding Non-coding

Protein-coding Protein-coding

Pseudogene Pseudogene

Sequence content Sequence content

CCDS CCDS

Ensembl Ensembl

RefSeq RefSeq

RefSeqGene RefSeqGene

Status Status

Current Current

Chromosome locations Chromosome locations

more...

[Clear all](#)

[Show additional filters](#)

Did you mean pp2a as a gene symbol?
Search Gene for [pp2a](#) as a symbol

Send to: [Manage Filters](#)

Results by taxon

Taxonomic Groups (List)

- eukaryotes (18126)
 - animals (7221)
 - chordates (5444)
 - arthropods (1335)
 - more... (441)
 - green plants (6290)
 - land plants (5961)
 - more... (329)
 - fungi (2516)
 - ascomycetes (1902)
 - more... (614)
 - apicomplexans (470)
 - more... (285)
 - oomycetes (253)
 - more... (253)
 - kinetoplastids (209)
 - more... (209)
 - Entamoeba (136)
 - more... (136)
 - cellular slime molds (80)
 - more... (616)
 - bacteria (425)
 - actinobacteria (133)
 - more... (124)
 - proteobacteria (124)
 - more... (109)
 - firmicutes (109)
 - more... (59)
 - archaea (55)
 - more... (41)
 - viruses (41)

Name/Gene ID	Description	Location	Aliases	MIM
<input checked="" type="checkbox"/> PP2A ID: 84333	serine/threonine protein phosphatase 2A [Arabidopsis thaliana (thale cress)]	Chromosome 1, NC_003070.9 (26348721..26350697, complement)	AT1G69960, F20P5.30, F20P5_30, TYPE 2A SERINE/THREONINE PROTEIN PHOSPHATASE, serine/threonine protein phosphatase 2A	
<input checked="" type="checkbox"/> pp2A ID: 3878393	protein phosphatase 2A-like [Neurospora crassa ORF4A]	Chromosome IV, NC_026504.1 (1637071..1639185, complement)	NCU06563	
<input checked="" type="checkbox"/> PP2A ID: 9680973	protein phosphatase 2A regulatory subunit [Micromonas pusilla COMP1545]		MICPUCDRAFT_30915	
<input checked="" type="checkbox"/> PPP2R4 ID: 5524	protein phosphatase 2 regulatory subunit 4 [Homo sapiens (human)]	Chromosome 9, NC_000009.12 (129110945..129148946)	PP2A, PR53, PTPA	600756
<input checked="" type="checkbox"/> Ppp2ca ID: 19052	protein phosphatase 2 (formerly 2A), catalytic subunit, alpha isozyme [Mus musculus (house mouse)]	Chromosome 11, NC_000077.6 (52098824..52122749)	PP2A, R75363	
<input checked="" type="checkbox"/> mts ID: 45059	microtubule star [Drosophila melanogaster]	Chromosome 2L, NT_033779.4	Dmel_CG7109_5569, CG7109_DmPp2A-2BD, DmelNCG7109_ER2-6, MTS/PP2A, Mts, PP2, PP2A, PP2A-2BD, PP2A-C, PP2A-D, PP2A-MTS, PP2A-ER2, PP2A-ER2-6, PP2a, PP2a-2BD, PP2a	

Find related data

Database: Select

Find items

Search details

pp2a[All Fields] AND alive[prop]

Search

See more...

Search NCBI

Gene sources
Genomic

Categories
Alternatively spliced
Annotated genes
 Protein-coding

Sequence content
 RefSeq

Status
 Current

Chromosome locations
more...

[Clear all](#)

[Show additional filters](#)

Tabular ▾ 20 per page ▾ Sort by Relevance ▾

Send to: ▾

[Hide sidebar >>](#)

Search results

Items: 1 to 20 of 67

Filters activated: Protein-coding, RefSeq. [Clear all](#) to show 57 items.

Showing Current items.

Name/Gene ID	Description	Location	Aliases
LOC100201603 ID: 100201603	serine/threonine-protein phosphatase 2A catalytic subunit beta isoform [<i>Hydra vulgaris</i>]		NEMVEDRAFT_v1g177129
NEMVEDRAFT_v1g177129 ID: 5522251	hypothetical protein [<i>Nematostella vectensis</i> (starlet sea anemone)]		NEMVEDRAFT_v1g175766
NEMVEDRAFT_v1g175766 ID: 5501708	hypothetical protein [<i>Nematostella vectensis</i> (starlet sea anemone)]		NEMVEDRAFT_v1g220857
NEMVEDRAFT_v1g220857 ID: 5500969	hypothetical protein [<i>Nematostella vectensis</i> (starlet sea anemone)]		NEMVEDRAFT_v1g195622
NEMVEDRAFT_v1g195622 ID: 5500404	hypothetical protein [<i>Nematostella vectensis</i> (starlet sea anemone)]		
LOC107353620 ID: 107353620	serine/threonine-protein phosphatase 2A activator-like [<i>Acropora digitifera</i>]		
LOC107352888 ID: 107352888	serine/threonine-protein phosphatase 2A catalytic subunit beta isoform [<i>Acropora digitifera</i>]		
LOC107349165 ID: 107349165	protein phosphatase 1H-like [<i>Acropora digitifera</i>]		

Find < Prev Page 1 of 3 Next > Last >>

Filters: [Manage Filters](#)

Results by taxon

Taxonomic Groups [List]

- cnidarians (57)
 - Anthozoa (39)
 - sea anemones (21)
 - stony corals (18)
 - hydromedusae (18)

Find related data

Database: [Select](#)

[Find aliases](#)

Search details

```
(pp2a[All Fields] AND "animals"[pongr] NOT "arthropods"
 [pongr] NOT "chordates"[pongr]) AND "cnidarians"[pongr] AND
 ("genotype protein coding"[Properties] AND "srcdb refseq"
 [Properties] AND alive[prop]))
```

[Search](#)

[See more...](#)

Search NCBI

LOC107353620 serine/threonine-protein phosphatase 2A activator-like [Acropora digitifera]

Gene ID: 107353620, updated on 16-Apr-2016

Summary

Gene symbol LOC107353620
 Gene description serine/threonine-protein phosphatase 2A activator-like
 Gene type protein coding
 RefSeq status MODEL
 Organism *Acropora digitifera*
 Lineage Eukaryota; Metazoa; Cnidaria; Anthozoa; Hexacorallia; Scleractinia; Astrocoeniina; Acroporidae; Acropora

Genomic context

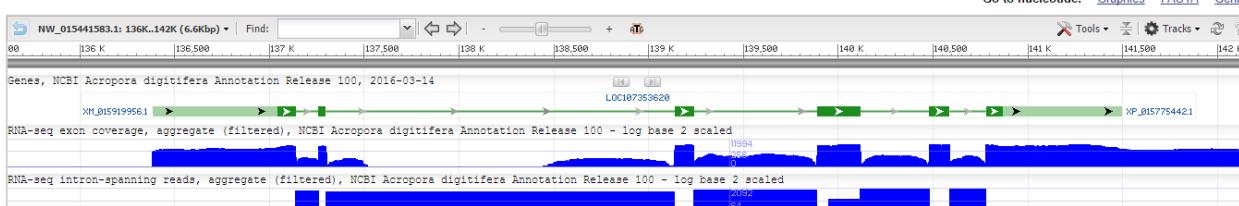
Location: chromosome: Un
 Exon count: 6

Annotation release	Status	Assembly	Chr	Location
100	current	Adig_1.1 (GCF_00022465.1)	Unplaced Scaffold	NW_015441583.1 (136387..141496)

Genomic regions, transcripts, and products

Genomic Sequence: NW_015441583.1 Unplaced Scaffold Reference Adig_1.1 Primary Assembly

Go to reference sequence details
 Go to nucleotide: [Graphics](#) [FASTA](#) [GenBank](#)



The figure shows a genomic track for LOC107353620 on an unplaced scaffold. The top part displays the gene structure with exons as green arrows and introns as black lines. The bottom part shows RNA-seq coverage as blue bars, with specific peaks labeled at positions 136,500, 137,500, 138,500, 139,500, 140,500, and 141,500. The RNA-seq data is from NCBI Acropora digitifera Annotation Release 100, dated 2016-03-14.

Table of contents

- Summary
- Genomic context
- Genomic regions, transcripts, and products
- Bibliography
- General protein information
- NCBI Reference Sequences (RefSeq)
- Related sequences

Genome Browsers

- Genome Data Viewer

Related information

- BioProjects
- Conserved Domains
- Gene neighbors
- Nucleotide
- Protein
- RefSeq Proteins
- RefSeq RNAs
- Taxonomy

General information

- About Gene
- FAQ
- FTP site
- Help
- My NCBI help
- NCBI Handbook
- Statistics

Related sites

- BLAST

Search NCBI

ⓘ NCBI is phasing out sequence GI numbers in September 2016. Please use accession.version! [Read more...](#)

GenBank ▾ Send: ▾ Change region shown

PREDICTED: Acropora digitifera serine/threonine-protein phosphatase 2A activator-like (LOC107353620), mRNA

NCBI Reference Sequence: XM_015919956.1

[FASTA](#) [Graphics](#)

Go to: ▾

LOCUS	XM_015919956	1942 bp	mRNA	linear	INV 14-MAR-2016
DEFINITION	PREDICTED: Acropora digitifera serine/threonine-protein phosphatase 2A activator-like (LOC107353620), mRNA.				
ACCESSION	XM_015919956				
VERSION	XM_015919956.1 GI:1005477360				
DBLINK	BioProject: PRJNA314803				
KEYWORDS	RefSeq.				
SOURCE	Acropora digitifera				
ORGANISM	Acropora digitifera				
Eukaryota; Metazoa; Cnidaria; Anthozoa; Hexacorallia; Scleractinia; Astrocoeniina; Acroporidae; Acropora.					
COMMENT	<p>MODEL REFSEQ: This record is predicted by automated computational analysis. This record is derived from a genomic sequence (NW_015441583.1) annotated using gene prediction method: Gnomon, supported by mRNA and EST evidence.</p> <p>Also see:</p> <p>Documentation of NCBI's Annotation Process</p> <pre>##Genome-Annotation-Data-START## Annotation Provider :: NCBI Annotation Status :: Full annotation Annotation Version :: Acropora digitifera Annotation Release 100 Annotation Pipeline :: NCBI eukaryotic genome annotation pipeline Annotation Software Version :: 6.5 Annotation Method :: Best-placed RefSeq; Gnomon Features Annotated :: Gene; mRNA; CDS; ncRNA ##Genome-Annotation-Data-END##</pre>				
FEATURES	Location/Qualifiers				
source	1..1942 /organism="Acropora digitifera" /mol_type="mRNA"				

Customize view

Analyze this sequence

Run BLAST

Pick Primers

Highlight Sequence Features

Find in this Sequence

Reference sequence information

RefSeq protein product
See the reference protein sequence for PREDICTED: serine/threonine-protein phosphatase 2A activator-like (XP_015775442.1).

More about the gene LOC107353620

LOC107353620 gene

Related information

Annotated Genomic

BioProject

Gene

Protein

Taxonomy

Recent activity

[Turn Off](#) [Clear](#)

Search NCBI

```

FEATURES          Location/Qualifiers
source           1..1942
/organism="Acropora digitifera"
/mol_type="mRNA"
/db_xref="taxon:70779"
/chromosome="Unknown"
/country="Japan:Okinawa, Kunigami, Oku"
1..1942
/gene="LOC107353620"
/note="Derived by automated computational analysis using
gene prediction method: Gnomon. Supporting evidence
includes similarity to: 6 mRNAs, 3 ESTs, 25 Proteins, and
100% coverage of the annotated genomic feature by RNAseq
alignments, including 25 samples with support for all
annotated introns"
/db_xref="GeneID:107353620"
659..1315
/gene="LOC107353620"
/codon_start=1
/product="serine/threonine-protein phosphatase 2A
activator-like"
/protein_id="XP_015775442.1"
/db_xref="GI:1005477361"
/db_xref="GeneID:107353620"
"MPLQQSVHSLVQPLLDPDKFLGAAIELTAYLKDAFGNKTRIDYGT
GHEASFAAFLCCLFKLRLVDQSDCAAIVFKVFQRYLELMRRLQLTYRMEPAGSQGVWG
LDDFQFLPIWGSQALIGHTSLEPQHFTCEKNVEEHHNKMFLGCIRFINQMKRGPF
EHSNTLWGSISSVKTWEKVNSGLMKMYKAEVLSKFPVIQHFVFGTLMSIKEGETFKKPL
"
ORIGIN
1 tcttgaggta cttgtgtctt cccagcattg actgaccctg gctgggtgaac aatgtttttt
61 tttttttttt tttttttcat ttgtttctaa gggtaggaga aactactatg aaaggctggg
121 aaaaatcaatt ttatccata cattaacttg tcactatata agtatgcctt tcactataaa
181 ctgaaaaatt aatgtgtatt ctttttaattc cattttctt tgttttttac attttgc当地
241 gtcctcaaaa agggtgatat agcaaacact tttcttcattc acagcttgc cacatcttg
301 gtttattgtga ctggccaaa agataggaaa atttttctg ctccaggag aaatataata
361 ttgcttgatt ctcagaatgg cctgcttata atctgagaaaa ttgatcattt ctggtaatta
421 aggtattgcc ttaatttttatt tgaatggca attggatttt gcttgcattc gcttattaa
481 aaagccatga attagaagag tccacatcat ttgggtgcag ttaagtaatt tccatagttt

```

Send: ▾ Change destination settings

Complete Record
 Coding Sequences
 Gene Features

Choose Destination

File Clipboard
 Collections Analysis Tool

Pick Primers

Send: ▾ Change destination settings

Complete Record
 Coding Sequences
 Gene Features

Download features.

Format

FASTA Nucleotide ▾

Create File

Search NCBI

sequence.txt

```
1 >loc|XM_015919956.1_cds_XP_015775442.1_1 [gene=LOC107353620] [protein=serine/threonine-protein phosphatase 2A activator-like] [protein_id=XP_015775442.1] [location=659..1315]
2 ATGCCCTCAACAGAGTGTCACTCGTTAGTCAACCTCTTACCCAGACAAGTTCTGGGCTGCCA
3 TTGAATTAAC TGCA TTTAAAGATGCC TTTGGAA AAAACAAGAACATAGACTATGGAACAGGT CATGA
4 AGCTTCCCTTGCTGCATTCTTTGGTTATTCAAGGCTCAGAGTATTGGACCAAAGTGACTGTGCTGCT
5 ATTGTGTTCAAGGTTTTAGAGGTATTAGAACTGATGAGACGATTG CAGCTCACTTACAGAA TGGAAC
6 CAGCTGGCAGTCAGGGTGTGGGGCTGGATGATTTCAGTTCTCCCTTCATTGGGGAA GTGCTCA
7 GCTGATAGGCCATACAAGTCTAGAGGCCACAGCACTTCACTTGTAAGAAAAACGTAAGAGGGACATATAAC
8 AAGTACATGTTCTGGGCTGCATCGTTTATAAACCAAATGAAAAGAGGACCC TTG CAGAACATTCCA
9 ACACTTGTGGGAATAAGCTGTAAAACATGGAAAAAGTAAACTCTGGTTGATGAAAATGTATAA
10 AGCTGAGGTTCTATCCAAGTCCCAGTCATT CAGCATTTGTTGGTACATTAATGTCTATAAAGAA
11 GGAGAAACGTTAAAAAACCCCTGTAA
12
13
```

Search NCBI

Species Summary ▾ 20 per page ▾ Sort by Default order ▾

Molecule types Selected: 3

Items: 1 to 20 of 736

1. [Caenorhabditis elegans Probable serine/threonine-protein phosphatase \(paa-1\). partial mRNA](#)
1,773 bp linear mRNA
Accession: NM_065761.4 GI: 392894997
GenBank FASTA Graphics

2. [Trichinella spiralis serine/threonine-protein phosphatase PP2A regulator, mRNA. partial cds](#)
159 bp linear mRNA
Accession: XM_003366924.1 GI: 339263803
GenBank FASTA Graphics

3. [Trichinella spiralis protein phosphatase PP2A \(Tsp_14159\) mRNA. partial cds](#)
394 bp linear mRNA
Accession: XM_003368715.1 GI: 339259723
GenBank FASTA Graphics

4. [Loa loa protein phosphatase PP2A regulatory subunit \(LOAG_00611\) mRNA. complete cds](#)
1,974 bp linear mRNA
Accession: XM_003136151.1 GI: 312066287
GenBank FASTA Graphics

Send to: ▾ **Filters:** [Manage Filters](#)

Choose Destination

File Clipboard Collections

Download 3 items.

Format: **FASTA**

Sort by: **Default order**

[Create File](#)

Find related data

Database: **Select**

[Find items](#)

Groups [List]

- Onion (6)
- Groups (26)
- Leucoscytidae (155)
- Leucostoma (58)
- Leucostomatidae (52)
- Leucostomatidae (22)
- Leucostomatidae (21)
- Leucostomatidae (15)
- Leucostomatidae (3)
- Leucostomatidae (265)
- Leucostomatidae (225)
- Trichuridae (40)
- Spirurida (96)
 - Loa (45)
 - Brugia (32)
 - Wuchereria (19)
- Ascaridida (36)
 - Toxocaridae (19)
 - Ascarididae (17)
 - Diplogasterida (12)
 - Tylenchida (1)

Search NCBI

Courier New | 11 | B 3 total sequences

Mode: Select / Slide Selection: 0 Sequence Mask: None
 Position: Numbering Mask: None Start ruler at: 1

Scroll speed slow ↴ fast

```
gi|339263803|ref|XM_003366924| AGCCCGTTGGAAAGGAATGCGAACGAGACCCCTGGTTGGGAGAAAGCAGTCGAATCATTGAGAACCTTAAAGTTGACAAACCTGGGATCACTGATTTAGTAAGTTAGATCCCTGTGGTGTCAGTRAGCGGAGTTCCAATAGGAAATGAA
gi|339259723|ref|XM_003368715| CCGATGTAATAATCATGCGTTGAATTAAATCATGATGGTAATTGCTGGCTACCGGTGATAAAGGCCGACGCATTGTAATATTCAACGATCARACCAATAAGCTGGTGAATGGTCATCGAAGTCCCGAATACAACGTTAGCACATTCAGAGCTCA
gi|312066287|ref|XM_003136151| ATGGAGTTGGAACATCGATTTAGGGGGTTGGGATGGGTGGTTAGTGGGGAGGGAGTCACCGGATCAGGACGAGTTAGTACRAGTTGAGAGGGAGGTGACCCAGGTCAAGGATGAGGTAGTGCACAGGGAGGAGGAGGAGTCAGG
```

Courier New | 11 | B 3 total sequences

Mode: Select / Slide Selection: null Sequence Mask: None
 Position: Numbering Mask: None Start ruler at: 1

Scroll speed slow ↴ fast

```
gi|339263803|ref|XM_003366924| SPLERIATVEETVVWEKAVESLRTLIVDKPWHDLEVKLDPVVGQLAAVARK*
gi|339259723|ref|XM_003368715| PM*YHALNLLIMMVNCWLEVIKA DAL*YENVVIKAISW*MVIEVLNTKCLHSRVVMNQNLI**NL*K*KKK*IVFDG*KGKT*LIFYFPLMIKQ*NYGK*LNVISGLMAVGICCTNTALPGY
gi|312066287|ref|XM_003136151| MEGFTIDLGGLGVGLVGRRESTESGRVSTVEEEGGDPGQDEVVHWERGGIQRVRTRLGYVVIMASLQAHEDTDNNLYPIAILIDELRNEDVQLRINSIRKLSTIALGVERTRGELIQ
```

Nucleotide Search

NCBI Resources How To Sign in to NCBI

Search NCBI databases

ambn Help

Results found in 25 databases for "ambn"

Literature			Genes		
Books	0	books and reports	EST	0	expressed sequence tag sequences
MeSH	5	ontology used for PubMed indexing	Gene	110	collected information about gene loci
NLM Catalog	0	books, journals and more in the NLM Collections	GEO DataSets	0	functional genomics studies
PubMed	190	scientific & medical abstracts/citations	GEO Profiles	2,920	gene expression and molecular abundance profiles
PubMed Central	233	full-text journal articles	HomoloGene	1	homologous gene sets for selected organisms
Health			PopSet	5	sequence sets from phylogenetic and population studies
ClinVar	10	human variations of clinical significance	UniGene	7	clusters of expressed transcripts
dbGaP	2	genotype/phenotype interaction studies	Proteins		
GTR	1	genetic testing registry	Conserved Domains	2	conserved protein domains
MedGen	1	medical genetics literature and links	Protein	3,160	protein sequences
OMIM	8	online mendelian inheritance in man	Protein Clusters	0	sequence similarity-based protein clusters
PubMed Health	0	clinical effectiveness, disease and drug reports	Structure	0	experimentally-determined biomolecular structures
Genomes			Chemicals		
Assembly	2	genome assembly information	BioSystems	39	molecular pathways with links to genes, proteins and chemicals
BioProject	0	biological projects providing data to NCBI	PubChem BioAssay	0	bioactivity screening studies
BioSample	0	descriptions of biological source materials	PubChem Compound	1	chemical information with structures, information and links
Clone	1,046	genomic and cDNA clones	PubChem Substance	91	deposited substance and chemical information
dbVar	110	genome structural variation studies			
Epigenomics	0	epigenomic studies and display tools			
Genome	3	genome sequencing projects by organism			
GSS	0	genome survey sequences			
Nucleotide	473	DNA and RNA sequences			
Probe	185	sequence-based probes and primers			
SNP	3,461	short genetic variations			
SRA	0	high-throughput DNA and RNA sequence read archive			
Taxonomy	0	taxonomic classification and nomenclature catalog			

You are here: NCBI > GQuery Write to the Help Desk

UniProt

UniProtKB ambn

Advanced Search

BLAST Align Retrieve/ID mapping Help Contact

About UniProtKB Basket

UniProtKB results

Filter by:

- Reviewed (5) Swiss-Prot
- Unreviewed (115) TrEMBL

Popular organisms

- Human (4)
- Mouse (3)
- Bovine (2)
- Rat (2)
- Pig (1)

Other organisms

Search terms

Filter "ambn" as: protein name (37)

View by

- Taxonomy
- Keywords
- Gene Ontology
- Enzyme class
- Pathway

UniRef

Your results in sequence clusters with

Entry	Entry name	Protein names	Gene names	Organism	Length
Q9NP70	AMBN_HUMAN	Ameloblastin	AMBN	Homo sapiens (Human)	447
Q55189	AMBN_MOUSE	Ameloblastin	Ambn	Mus musculus (Mouse)	407
Q28989	AMBN_PIG	Ameloblastin	AMBN	Sus scrofa (Pig)	421
Q62840	AMBN_RAT	Ameloblastin	Ambn	Rattus norvegicus (Rat)	422
Q5M8P3	Q5M8P3_MOUSE	Ambn protein	Ambn mCG_119077	Mus musculus (Mouse)	422
Q9XSX7	AMBN_BOVIN	Ameloblastin	AMBN	Bos taurus (Bovine)	392
Q546D7	Q546D7_HUMAN	Ameloblastin	AMBN	Homo sapiens (Human)	447
Q3B861	Q3B861_HUMAN	AMBN protein	AMBN	Homo sapiens (Human)	446
Q811C6	Q811C6_CAVPO	Ameloblastin	ambn AMBN	Cavia porcellus (Guinea pig)	423
Q811C5	Q811C5_CAVPO	Ameloblastin	ambn AMBN	Cavia porcellus (Guinea pig)	407
B1ACP5	B1ACP5_CAPMR	AMBN	AMBN	Caperea marginata (Pygmy right whale) (Balaena marginata)	155
B1ACP4	B1ACP4_BALAC	AMBN	AMBN	Balaenoptera acutorostrata (Common minke whale) (Balaena rostrata)	156
B1ACQ4	B1ACQ4_PECTA	AMBN	AMBN	Pecari tajacu (Collared peccary) (Tayassu tajacu)	155
B1ACP7	B1ACP7_DELLE	AMBN	AMBN	Delphinapterus leucas (Beluga whale)	155
B1ACQ7	B1ACQ7_TAPIN	AMBN	AMBN	Tapirus indicus (Asiatic tapir) (Malayan tapir)	155
B1ACQ5	B1ACQ5_CAMDR	AMBN	AMBN	Camelus dromedarius (Dromedary) (Arabian camel)	150

UniProt

PTM / Processingⁱ

Molecule processing

Feature key	Position(s)	Length	Description	Graphical view	Feature identifier	Actions
Signal peptide ⁱ	1 – 26	26	Sequence analysis			 Add  BLAST
Chain ⁱ	27 – 447	421	Ameloblastin		PRO_0000001192	 Add  BLAST

Amino acid modifications

Feature key	Position(s)	Length	Description	Graphical view	Feature identifier	Actions
Modified residue ⁱ	37 – 37	1	Hydroxyproline By similarity			
Modified residue ⁱ	43 – 43	1	Phosphoserine By similarity			
Glycosylation ⁱ	112 – 112	1	O-linked (GalNAc...) By similarity			

Keywords - PTMⁱ

Glycoprotein, Hydroxylation, Phosphoprotein

Proteomic databases

PaxDb ⁱ	Q9NP70.
PRIDE ⁱ	Q9NP70.

PTM databases

PhosphoSite ⁱ	Q9NP70.
--------------------------	---------

Expressionⁱ

Tissue specificityⁱ

Ameloblast-specific. Located at the Tomes processes of secretory ameloblasts and in the sheath space between rod-interrod enamel.

Gene expression databases

UniProt

Interactionⁱ

GO - Molecular functionⁱ

- growth factor activity 

Protein-protein interaction databases

BioGrid ⁱ	106756. 2 interactions.
STRING ⁱ	9606.ENSP00000313809.

Structureⁱ

3D structure databases

ProteinModelPortal ⁱ	Q9NP70.
ModBase ⁱ	Search...
MobiDB ⁱ	Search...

Family & Domainsⁱ

Domains and Repeats

Feature key	Position(s)	Length	Description	Graphical view	Feature identifier	Actions
Repeat ⁱ	189 – 201	13	1			 Add  BLAST
Repeat ⁱ	202 – 214	13	2			 Add  BLAST

Sequence similaritiesⁱ

Belongs to the ameloblastin family. 

Keywords - Domainⁱ

Repeat, Signal

Hands-On 2

1. Escolher dois genes:
 - Acetylcholinesterase (ACHE)
 - Ameloblastin (AMBN)
 - Enamelin (ENAM)
 - Hemopexin (HPX)
2. Ir à página do NCBI e descarregar sequencias de 10 espécies

NCBI – BLAST (Basic Local Alignment Serach Tool)

BLAST

Home Recent Results Saved Strategies Help

NCBI/ BLAST Home

BLAST finds regions of similarity between biological sequences. [more...](#)

New Aligning Multiple Protein Sequences? Try the COBALT Multiple Alignment Tool. [Go](#)

BLAST Assembled Genomes

Choose a species genome to search, or [list all genomic BLAST databases](#).

<input type="checkbox"/> Human	<input type="checkbox"/> Oryza sativa	<input type="checkbox"/> Gallus gallus
<input type="checkbox"/> Mouse	<input type="checkbox"/> Bos taurus	<input type="checkbox"/> Pan troglodytes
<input type="checkbox"/> Rat	<input type="checkbox"/> Danio rerio	<input type="checkbox"/> Microbes
<input type="checkbox"/> Arabidopsis thaliana	<input type="checkbox"/> Drosophila melanogaster	<input type="checkbox"/> Apis mellifera

Basic BLAST

Choose a BLAST program to run.

nucleotide blast	Search a nucleotide database using a nucleotide query <i>Algorithms: blash, megablast, discontiguous megablast</i>
protein blast	Search protein database using a protein query <i>Algorithms: blasp, psi-blast, phi-blast</i>
blastx	Search protein database using a translated nucleotide query
tblastn	Search translated nucleotide database using a protein query
tblastx	Search translated nucleotide database using a translated nucleotide query

Specialized BLAST

Choose a type of specialized search (or database name in parentheses.)

Make specific primers with [Primer-BLAST](#)

News

[BLAST 2.2.23 release](#)
A new version of the stand-alone applications is available.
Mon, 22 Mar 2010 15:00:00 EST
[More BLAST news...](#)

Tip of the Day

[How to do Batch BLAST jobs.](#)
BLAST makes it easy to examine a large group of potential gene candidates.
[More tips...](#)

Blast Search

BLAST Basic Local Alignment Search Tool

Home Recent Results Saved Strategies Help

NCBI/ BLAST/ blastn suite

blastn blastp blastx tblastn tblastx

Enter Query Sequence

Enter accession number, gi, or FASTA sequence Query subrange

Or, upload file

Job Title Enter a descriptive title for your BLAST search

Align two or more sequences

Choose Search Set

Database Human genomic + transcript Mouse genomic + transcript Others (nr etc.):
Nucleotide collection (nr/nt)

Organism Optional Exclude

Exclude Optional Models (XM/XP) Uncultured/environmental sample sequences

Entrez Query Optional Enter an Entrez query to limit search

Program Selection

Optimize for Highly similar sequences (megablast)
 More dissimilar sequences (discontiguous megablast)
 Somewhat similar sequences (blastn)
 Choose a BLAST algorithm

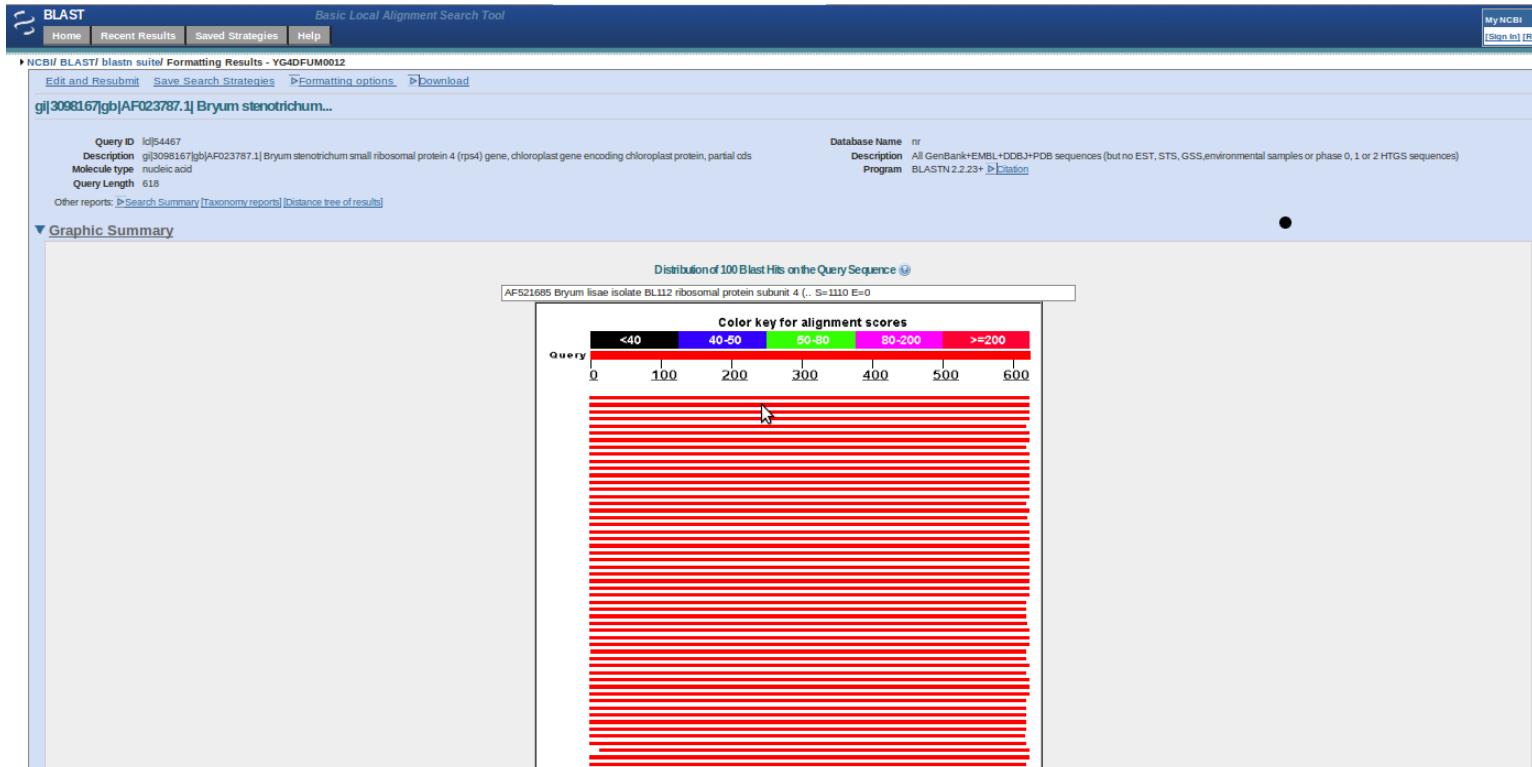
BLAST Search database Nucleotide collection (nr/nt) using Blastn (Optimize for somewhat similar sequences) Show results in a new window

Algorithm parameters Note: Parameter values that differ from the default are highlighted in yellow and marked with * sign

Blast methods

- blastn
- blastp
- blastx
- tblastn
- tblastx

NCBI – nucleotide BLAST result



NCBI – nucleotide BLAST result

BLAST Basic Local Alignment Search Tool

Home Recent Results Saved Strategies Help

NCBI/ BLAST! blasts suite! Formatting Results · YG4DFUM0012

Edit and Resubmit Save Search Strategies ▾Formatting options ▾Download

gi|3098167|gb|AF023787.1| Bryum stenorhizum...

Query ID: id|34467
Description: gi|3098167|gb|AF023787.1| Bryum stenorhizum small ribosomal protein 4 (rps4) gene, chloroplast gene encoding chloroplast protein, partial cds
Molecule type: nucleic acid
Query Length: 618

Database Name: nr
Description: All GenBank+EMBL+DDBJ+PDB sequences (but no EST, STS, GSS environmental samples or phase 0, 1 or 2 HTGS sequences)
Program: BLASTN 2.2.23+ ▾Citation

Other reports: ▾Search Summary [Taxonomy reports] [Distance tree of results]

► Graphic Summary

▼ Descriptions

Legend for links to other resources: U UniGene E GEO G Gene S Structure M Map Viewer

Sequences producing significant alignments:

Accession	Description	Max score	Total score	Query coverage	E value	Max ident	Links
AF023787.1	Bryum stenorhizum small ribosomal protein 4 (rps4) gene, chloroplast gene encoding chloroplast protein, partial cds	1115	1115	100%	0.0	100%	
AF21082.1	Bryum lisae isolate BL112 ribosomal protein subunit 4 (rps4) gene, partial cds; chloroplast gene for chloroplast product	1110	1110	100%	0.0	99%	
AF521080.1	Bryum arachnellae isolate BA107 ribosomal protein subunit 4 (rps4) gene, partial cds; chloroplast gene for chloroplast product	1110	1110	100%	0.0	99%	
AY082394.1	Bryum radicum small ribosomal subunit 4 (rps4) gene, partial cds; chloroplast gene for chloroplast product	1106	1106	100%	0.0	99%	
AY078333.1	Bryum pallescens small ribosomal subunit protein 4 (rps4) gene, partial cds; chloroplast gene for chloroplast product	1103	1103	99%	0.0	99%	
AF521689.1	Bryum pseudotriquetrum isolate BA116 ribosomal protein subunit 4 (rps4) gene, partial cds; chloroplast gene for chloroplast product	1101	1101	100%	0.0	99%	
AF023785.1	Bryum donarium small ribosomal protein 4 (rps4) gene, chloroplast gene encoding chloroplast protein, partial cds	1097	1097	100%	0.0	99%	
AY078329.1	Bryum purpureascens small ribosomal subunit protein 4 (rps4) gene, partial cds; chloroplast gene for chloroplast product	1094	1094	99%	0.0	99%	
AF521678.1	Bryum algivorum isolate BA105 ribosomal protein subunit 4 (rps4) gene, partial cds; chloroplast gene for chloroplast product	1092	1092	100%	0.0	99%	
AY090771.1	Rosularium albolimbatum isolate MDP423 small ribosomal protein subunit 4 (rps4) gene, partial cds; and tRNA-Ser ger	1092	1092	100%	0.0	99%	
AF521692.1	Haplodontium reticulatum isolate HR119 ribosomal protein subunit 4 (rps4) gene, partial cds; chloroplast gene for chloroplast product	1090	1090	100%	0.0	99%	
AF521673.1	Acidodontium heterodon isolate AH100 ribosomal protein subunit 4 (rps4) gene, partial cds; chloroplast gene for chloroplast product	1088	1088	100%	0.0	99%	
AY163087.1	Brachymenium preissianum ribosomal protein subunit 4 gene, partial cds; chloroplast gene for chloroplast product	1088	1088	100%	0.0	99%	
AF521694.1	Bryum c. caucasicum isolate MH121 ribosomal protein subunit 4 (rps4) gene, partial cds; chloroplast gene for chloroplast product	1083	1083	100%	0.0	98%	
AF521682.1	Bryum capillare isolate BC109 ribosomal protein subunit 4 (rps4) gene, partial cds; chloroplast gene for chloroplast product	1083	1083	100%	0.0	98%	
AY078320.1	Brachymenium acuminatum small ribosomal subunit protein 4 (rps4) gene, partial cds; chloroplast gene for chloroplast product	1081	1081	99%	0.0	99%	
DQ294323.1	Bryum bicolor small ribosomal subunit protein 4 (rps4) gene, partial cds; chloroplast	1079	1079	100%	0.0	98%	
AF521687.1	Bryum pachythecia isolate EP114 ribosomal protein subunit 4 (rps4) gene, partial cds; chloroplast gene for chloroplast product	1079	1079	99%	0.0	98%	
AY090773.1	Plagiothecium zeni isolate MDP207 small ribosomal protein subunit 4 gene, partial cds; and tRNA-Ser gene, partial cds	1079	1079	100%	0.0	98%	
AY163091.1	Bryum orthothecium ribosomal protein subunit 4 gene, partial cds; chloroplast gene for chloroplast product	1079	1079	100%	0.0	98%	
AY163090.1	Bryum coronatum ribosomal protein subunit 4 gene, partial cds; chloroplast gene for chloroplast product	1079	1079	100%	0.0	98%	
AY163086.1	Bryum clavatum ribosomal protein subunit 4 gene, partial cds; chloroplast gene for chloroplast product	1079	1079	100%	0.0	98%	
AY163086.1	Brachymenium philonotula ribosomal protein subunit 4 gene, partial cds; chloroplast gene for chloroplast product	1079	1079	100%	0.0	98%	
AY082593.1	Bryum nudale small ribosomal subunit 4 (rps4) gene, partial cds; chloroplast gene for chloroplast product	1079	1079	100%	0.0	98%	
AF21311.1	Mielichhoferia macrocarpa chloroplast partial rps4 gene for ribosomal protein, subunit 4	1079	1079	100%	0.0	98%	
AF023786.1	Anomobryum julaceum small ribosomal protein 4 (rps4) gene, chloroplast gene encoding chloroplast protein, partial cds	1077	1077	100%	0.0	98%	
AF521690.1	Bryum uliginosum isolate BU117 ribosomal protein subunit 4 (rps4) gene, partial cds; chloroplast gene for chloroplast product	1074	1074	100%	0.0	98%	
AF521676.1	Anomobryum conicum isolate AC103 ribosomal protein subunit 4 (rps4) gene, partial cds; chloroplast gene for chloroplast product	1074	1074	100%	0.0	98%	

E-values: 10^{-4} often considered good enough for an assumption of homology

Query	961	GGAGGTGCACAAGGCTCCCCTATGCCGGAGGCCAACCCAGACAATCTAGAAAACCCAGCT	1020
Sbjct	1062	GGAGGTGCACAAGGCTCCCCTATGCCGGAGGCCAACCCAGACAATCTAGAAAACCCAGCT	1121
Query	1021	TTCCTTACAGAGCTAGAACCTGCTCCCCACGCAGGGCTCCTTGCTCTCCCTAAGGATGAC	1080
Sbjct	1122	TTCCTTACAGAGCTAGAACCTGCTCCCCACGCAGGGCTCCTTGCTCTCCCTAAGGATGAC	1181
Query	1081	ATTCCCGGCCTGCCAAGGAGCCCTTCAGGAAAGATGAAGGGACTCCCCAGCGTCACCCCA	1140
Sbjct	1182	ATTCCCGGCCTGCCAAGGAGCCCTTCAGGAAAGATGAAGGGACTCCCCAGCGTCACCCCA	1241
Query	1141	GCAGCTGCTGACCCACTGATGACCCCTGAATTAGCTGATTTATAGGACCTACGATGCT	1200
Sbjct	1242	GCAGCTGCTGACCCACTGATGACCCCTGAATTAGCTGATTTATAGGACCTACGATGCT	1301
Query	1201	GACATGACCACATCCGTGGATTTCAGGAAAGCAACCATGGATACCCAGATGGCCCCA	1260
Sbjct	1302	GACATGACCACATCCGTGGATTTCAGGAAAGCAACCATGGATACCCAGATGGCCCCA	1361
Query	1261	AACTCTCTGCAAACATCCATGCCAGGAAACAAAGCCAGGAGGCCGAGATGATGCATGAC	1320
Sbjct	1362	AACTCTCTGCAAACATCCATGCCAGGAAACAAAGCCAGGAGGCCGAGATGATGCATGAC	1421
Query	1321	GCATGGCATTTCAGAGGCCCTG	1343
Sbjct	1422	GCATGGCATTTCAGAGGCCCTG	1444

Range 2: 633 to 671 [GenBank](#) [Graphics](#) ▼ Next Match ▲ Previous Match ⚡ First Match

Score	Expect	Identities	Gaps	Strand
53.6 bits(58)	0.011	35/39(90%)	0/39(0%)	Plus/Plus

Query	610	CTCCCAAGGATTGGATTTGCTGATCCACAAGGTTCAACA	648
Sbjct	633	CTCCCAAGGAGTAGATTTGCTGATCCACAAGGTTCCATCA	671

Range 3: 711 to 749 [GenBank](#) [Graphics](#) ▼ Next Match ▲ Previous Match ⚡ First Match

Score	Expect	Identities	Gaps	Strand
53.6 bits(58)	0.011	35/39(90%)	0/39(0%)	Plus/Plus

Query	532	CTCCCAAGGAGTAGATTTGCTGATCCACAAGGTTCCATCA	570
Sbjct	711	CTCCCAAGGATTGGATTTGCTGATCCACAAGGTTCAACA	749

Warnings

- Predictions methods can fail and sometimes accuracy is not available
- Prediction is always made of known issues
- Databases can contain incorrect data
- Avoid overvaloration of results

Hands-On 3

- 1) Descarregar sequencias da pasta hands_on_3
- 2) Fazer o “blast” para as sequencias
- 3) Identificar gene e espécies
- 4) Descarregar 3 top hits



Methods for multiple sequence alignments

- Web-based
- Local (software)
- Scripts

- Model
- Can be biased
- Quality can be checked (e.g.
<http://guidance.tau.ac.il/ver2/>)
- Parameters:
 - Gap penalties
 - Mismatch
 - Iterations
 - Guiding tree

MSA Algorithms

- ClustalW
- Muscle
- T-Coffee
- MAFT
- ClustalOmega
- Prank

MSA

FGFR1_10 -----
 FGFR1_12 -----
 FGFR1_15 MNSWKCLLFNAVLVITATLCTARPSPTLPE-----
 FGFR1_6 MNSWKCLLFNAVLVITATLCTARPSPTLPE-----
 FGFR1_8 MNSWKCLLFNAVLVITATLCTARPSPTLPE-----
 FGFR1_14 MNSWKCLLFNAVLVITATLCTARPSPTLPEAQCPNGAPVEVESFLVNPGLLQLRCRLRDLVQSINWLRDGVLQAESENTRIIGEEEVQDSVPADSGLYACVTSSPSGEDTTTFSVNVSDALPSSBDDDDDDSSSEEKETDNTKPNRMP-----
 FGFR1_4 MNSWKCLLFNAVLVITATLCTARPSPTLPEAQCPNGAPVEVESFLVNPGLLQLRCRLRDLVQSINWLRDGVLQAESENTRIIGEEEVQDSVPADSGLYACVTSSPSGEDTTTFSVNVSDALPSSBDDDDDDSSSEEKETDNTKPNRMP-----
 FGFR1 MNSWKCLLFNAVLVITATLCTARPSPTLPEAQCPNGAPVEVESFLVNPGLLQLRCRLRDLVQSINWLRDGVLQAESENTRIIGEEEVQDSVPADSGLYACVTSSPSGEDTTTFSVNVSDALPSSBDDDDDDSSSEEKETDNTKPNRMP-----
 FGFR1_11 -----
 FGFR1_13 -----
 FGFR1_7 MNSWKCLLFNAVLVITATLCTARPSPTLPE-----
 FGFR1_9 MNSWKCLLFNAVLVITATLCTARPSPTLPE-----
 FGFR1_17 MNSWKCLLFNAVLVITATLCTARPSPTLPE-----
 FGFR1_18 MNSWKCLLFNAVLVITATLCTARPSPTLPE-----
 FGFR1_2 MNSWKCLLFNAVLVITATLCTARPSPTLPEAQCPNGAPVEVESFLVNPGLLQLRCRLRDLVQSINWLRDGVLQAESENTRIIGEEEVQDSVPADSGLYACVTSSPSGEDTTTFSVNVSDALPSSBDDDDDDSSSEEKETDNTKPNRMP-----
 FGFR1_5 MNSWKCLLFNAVLVITATLCTARPSPTLPEAQCPNGAPVEVESFLVNPGLLQLRCRLRDLVQSINWLRDGVLQAESENTRIIGEEEVQDSVPADSGLYACVTSSPSGEDTTTFSVNVSDALPSSBDDDDDDSSSEEKETDNTKPNRMP-----
 FGFR1_3 MNSWKCLLFNAVLVITATLCTARPSPTLPEAQ-----
 FGFR1_16 MNSWKCLLFNAVLVITATLCTARPSPTLPEAQ-----
 ruler 1.....10.....20.....30.....40.....50.....60.....70.....80.....90.....100.....110.....120.....130.....140.....150

FGFR1_10 -----MEKKLHAVPAAKTVFKCPSSGTIPPTLRLWLNGKEEFPDHRIGGYKVRYATWSIIMDSVVPSDKGNYTCIVEENEWSINHTYLDVVERSPHRPILQAGLPANKTVALGSNVEFCKVSDPQPHIQWLKHIEVNNGSKI-----
 FGFR1_12 -----MEKKLHAVPAAKTVFKCPSSGTIPPTLRLWLNGKEEFPDHRIGGYKVRYATWSIIMDSVVPSDKGNYTCIVEENEWSINHTYLDVVERSPHRPILQAGLPANKTVALGSNVEFCKVSDPQPHIQWLKHIEVNNGSKI-----
 FGFR1_15 VAPIWTSPEKMEKKLHAVPAAKTVFKCPSSGTIPPTLRLWLNGKEEFPDHRIGGYKVRYATWSIIMDSVVPSDKGNYTCIVEENEWSINHTYLDVVERSPHRPILQAGLPANKTVALGSNVEFCKVSDPQPHIQWLKHIEVNNGSKI-----
 FGFR1_6 VAPIWTSPEKMEKKLHAVPAAKTVFKCPSSGTIPPTLRLWLNGKEEFPDHRIGGYKVRYATWSIIMDSVVPSDKGNYTCIVEENEWSINHTYLDVVERSPHRPILQAGLPANKTVALGSNVEFCKVSDPQPHIQWLKHIEVNNGSKI-----
 FGFR1_8 VAPIWTSPEKMEKKLHAVPAAKTVFKCPSSGTIPPTLRLWLNGKEEFPDHRIGGYKVRYATWSIIMDSVVPSDKGNYTCIVEENEWSINHTYLDVVERSPHRPILQAGLPANKTVALGSNVEFCKVSDPQPHIQWLKHIEVNNGSKI-----
 FGFR1_14 VAPIWTSPEKMEKKLHAVPAAKTVFKCPSSGTIPPTLRLWLNGKEEFPDHRIGGYKVRYATWSIIMDSVVPSDKGNYTCIVEENEWSINHTYLDVVERSPHRPILQAGLPANKTVALGSNVEFCKVSDPQPHIQWLKHIEVNNGSKI-----
 FGFR1_4 VAPIWTSPEKMEKKLHAVPAAKTVFKCPSSGTIPPTLRLWLNGKEEFPDHRIGGYKVRYATWSIIMDSVVPSDKGNYTCIVEENEWSINHTYLDVVERSPHRPILQAGLPANKTVALGSNVEFCKVSDPQPHIQWLKHIEVNNGSKI-----
 FGFR1 VAPIWTSPEKMEKKLHAVPAAKTVFKCPSSGTIPPTLRLWLNGKEEFPDHRIGGYKVRYATWSIIMDSVVPSDKGNYTCIVEENEWSINHTYLDVVERSPHRPILQAGLPANKTVALGSNVEFCKVSDPQPHIQWLKHIEVNNGSKI-----
 FGFR1_11 -----MEKKLHAVPAAKTVFKCPSSGTIPPTLRLWLNGKEEFPDHRIGGYKVRYATWSIIMDSVVPSDKGNYTCIVEENEWSINHTYLDVVERSPHRPILQAGLPANKTVALGSNVEFCKVSDPQPHIQWLKHIEVNNGSKI-----
 FGFR1_13 -----MEKKLHAVPAAKTVFKCPSSGTIPPTLRLWLNGKEEFPDHRIGGYKVRYATWSIIMDSVVPSDKGNYTCIVEENEWSINHTYLDVVERSPHRPILQAGLPANKTVALGSNVEFCKVSDPQPHIQWLKHIEVNNGSKI-----
 FGFR1_7 VAPIWTSPEKMEKKLHAVPAAKTVFKCPSSGTIPPTLRLWLNGKEEFPDHRIGGYKVRYATWSIIMDSVVPSDKGNYTCIVEENEWSINHTYLDVVERSPHRPILQAGLPANKTVALGSNVEFCKVSDPQPHIQWLKHIEVNNGSKI-----
 FGFR1_9 VAPIWTSPEKMEKKLHAVPAAKTVFKCPSSGTIPPTLRLWLNGKEEFPDHRIGGYKVRYATWSIIMDSVVPSDKGNYTCIVEENEWSINHTYLDVVERSPHRPILQAGLPANKTVALGSNVEFCKVSDPQPHIQWLKHIEVNNGSKI-----
 FGFR1_17 VAPIWTSPEKMEKKLHAVPAAKTVFKCPSSGTIPPTLRLWLNGKEEFPDHRIGGYKVRYATWSIIMDSVVPSDKGNYTCIVEENEWSINHTYLDVVERSPHRPILQAGLPANKTVALGSNVEFCKVSDPQPHIQWLKHIEVNNGSKI-----
 FGFR1_18 VAPIWTSPEKMEKKLHAVPAAKTVFKCPSSGTIPPTLRLWLNGKEEFPDHRIGGYKVRYATWSIIMDSVVPSDKGNYTCIVEENEWSINHTYLDVVERSPHRPILQAGLPANKTVALGSNVEFCKVSDPQPHIQWLKHIEVNNGSKI-----
 FGFR1_2 VAPIWTSPEKMEKKLHAVPAAKTVFKCPSSGTIPPTLRLWLNGKEEFPDHRIGGYKVRYATWSIIMDSVVPSDKGNYTCIVEENEWSINHTYLDVVERSPHRPILQAGLPANKTVALGSNVEFCKVSDPQPHIQWLKHIEVNNGSKI-----
 FGFR1_5 VAPIWTSPEKMEKKLHAVPAAKTVFKCPSSGTIPPTLRLWLNGKEEFPDHRIGGYKVRYATWSIIMDSVVPSDKGNYTCIVEENEWSINHTYLDVVERSPHRPILQAGLPANKTVALGSNVEFCKVSDPQPHIQWLKHIEVNNGSKI-----
 FGFR1_3 -----CPDLIAEKSCSAFPHSI-----
 FGFR1_16 -----PWGAIVEVBSLEVHPGDLQLRCRLEDDW-QSINWLRDGVLQAESENTRIITGEVEVQDSVPAASGLACVTSSPSGEDTTTFSVNVS-----ACPDLIAEKSCSAFPHSI-----
 ruler160.....170.....180.....190.....200.....210.....220.....230.....240.....250.....260.....270.....280.....290.....300

Hands-On 4

- 1) Requisitos: Seaview e/ou Mega6
- 2) Descarregar sequencias da pasta Hands_on_4
- 3) Importar ficheiro para o Mega
- 4) Alinhar usado o Muscle/ClustalW
- 5) Traduzir para aminoácidos
- 6) Repetir o passo 3
- 7) Repetir passos 3 a 6 usando o Seaview

Q & A



INTRODUÇÃO À BIOLOGIA MOLECULAR E BIOINFORMÁTICA: Bioinformatics 2

Lisboa, 1-3 Junho

João Machado

Bárbara Frazão



Phylogenetic trees

The tree of Life

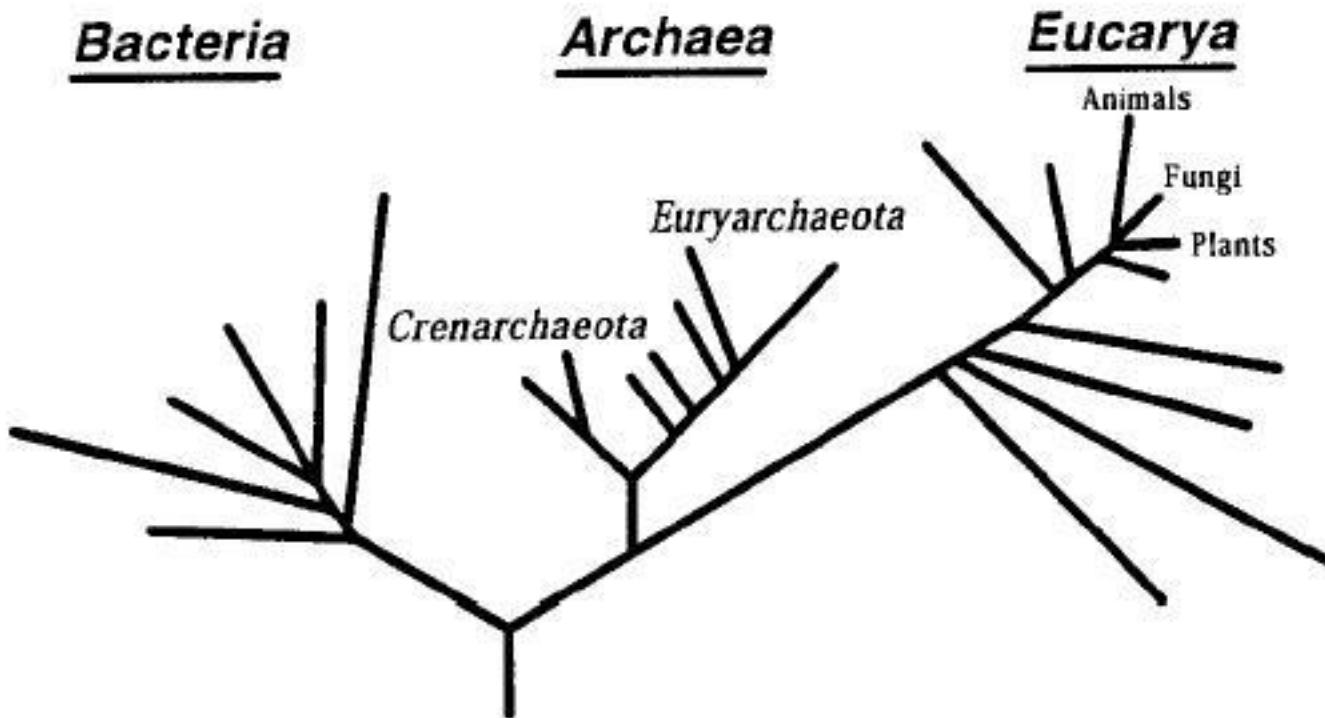
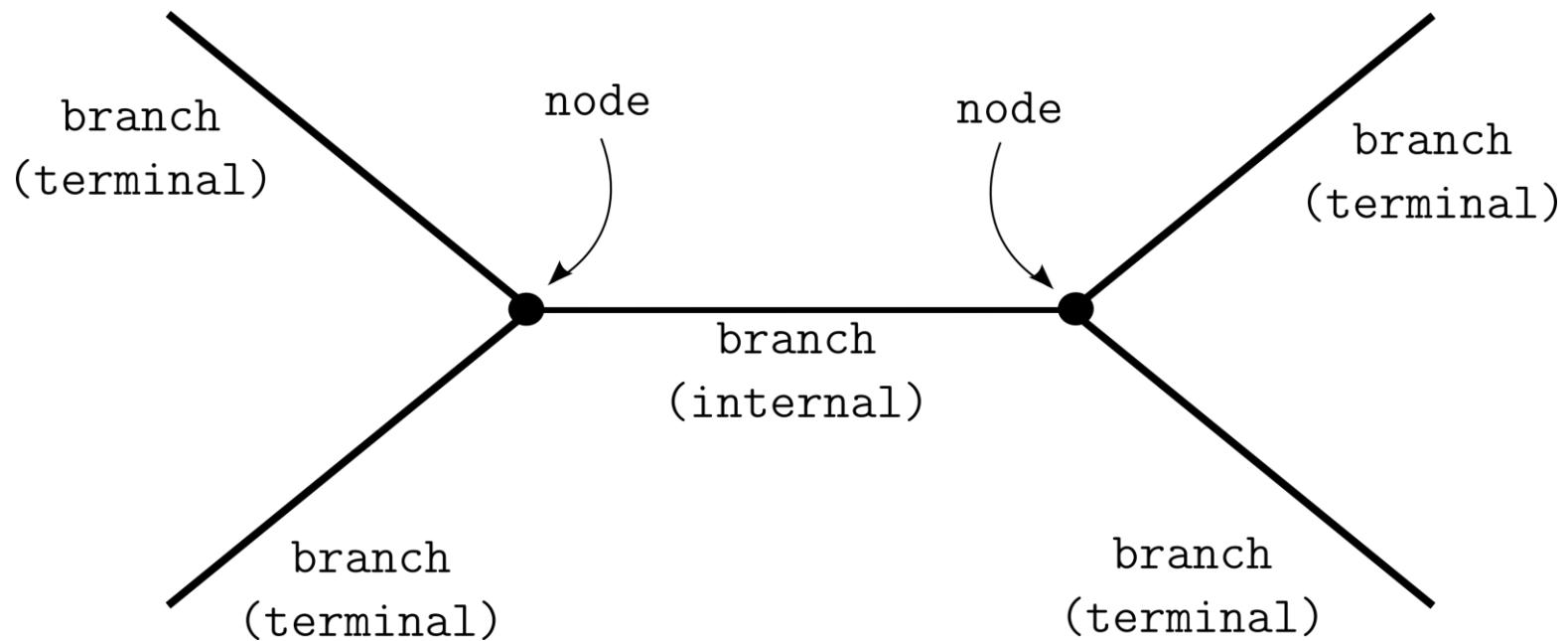
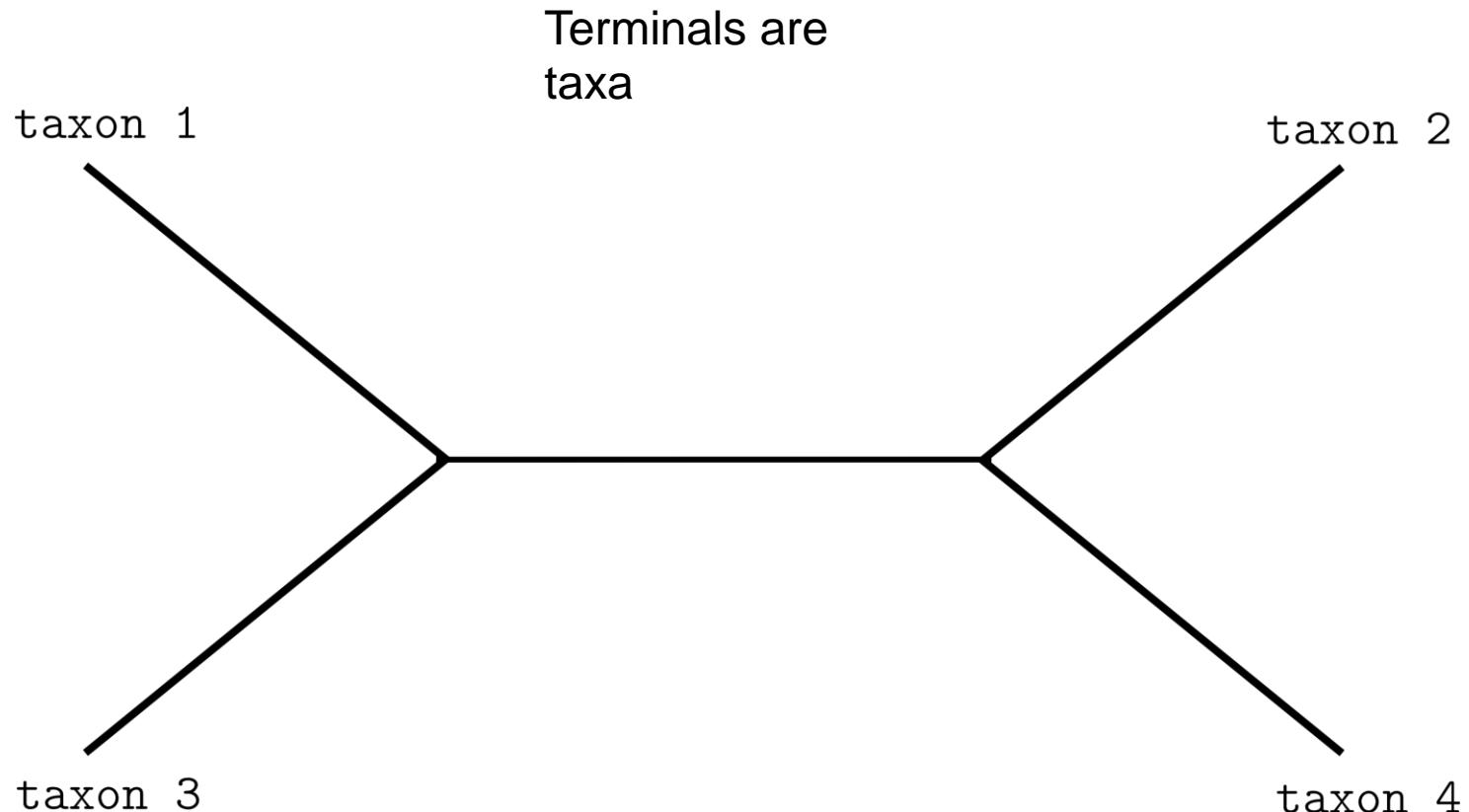


Fig. 1. The basal universal phylogenetic tree inferred from comparative analyses of rRNA sequences (4, 5). The root has been determined by using the paralogous gene couple EF-Tu/EFG (6).

Reading Trees



Reading Trees



Rooted / Unrooted Tree

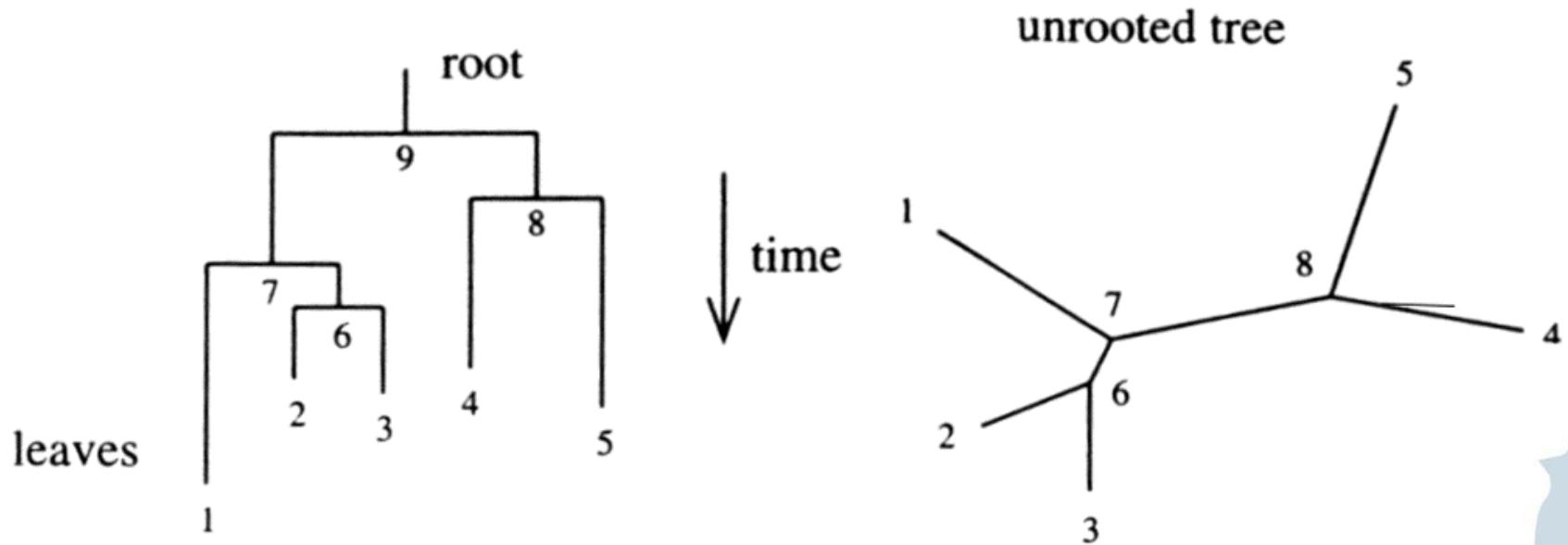
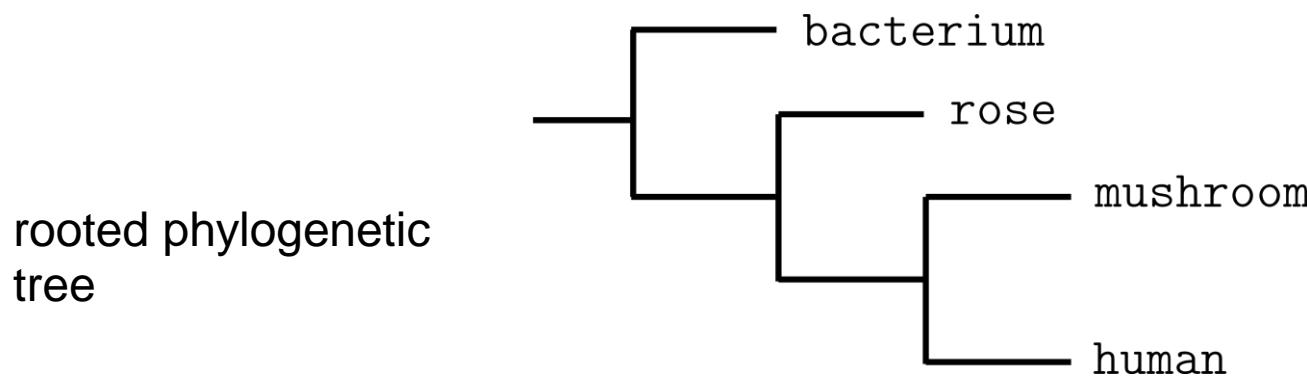
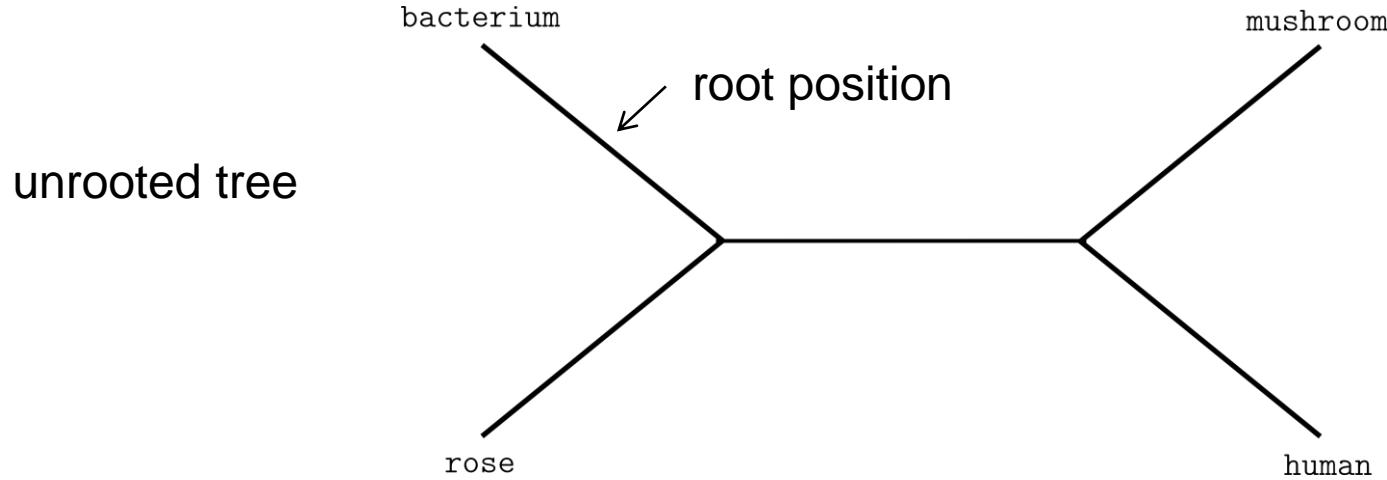
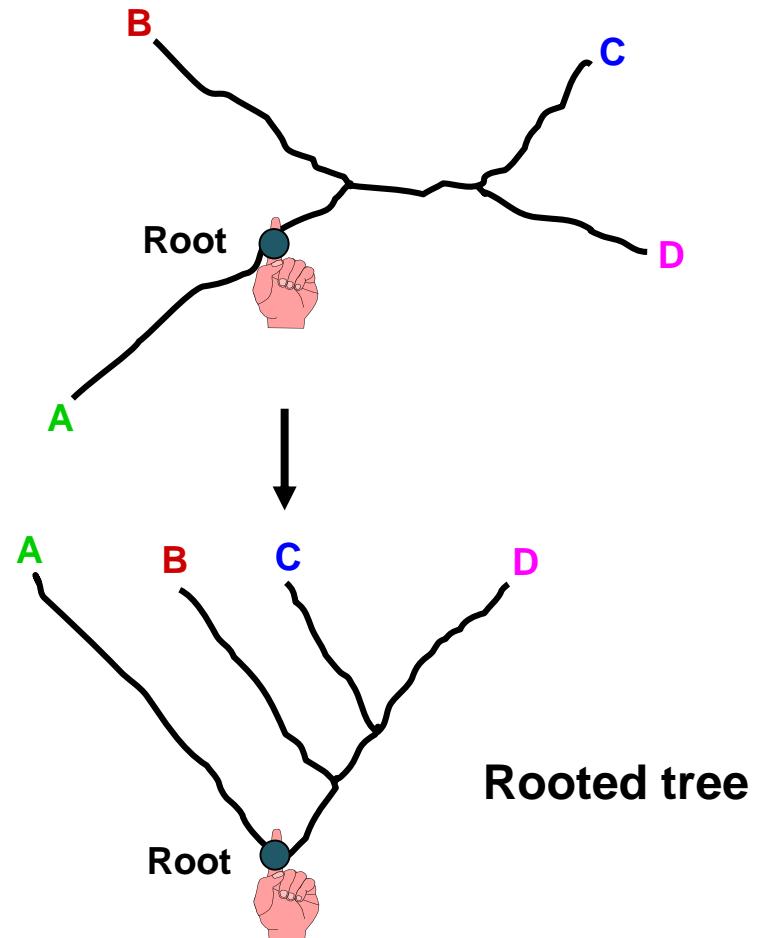


Figure 7.2 An example of a binary tree, showing the root and leaves, and the direction of evolutionary time (the most recent time being at the bottom of the figure). The corresponding unrooted tree is also shown; the direction of time here is undetermined.

Rooted / Unrooted Tree



Rooted / Unrooted Tree



Note that in this rooted tree, taxon A is no more closely related to taxon B than it is to C or D.

Counting Trees

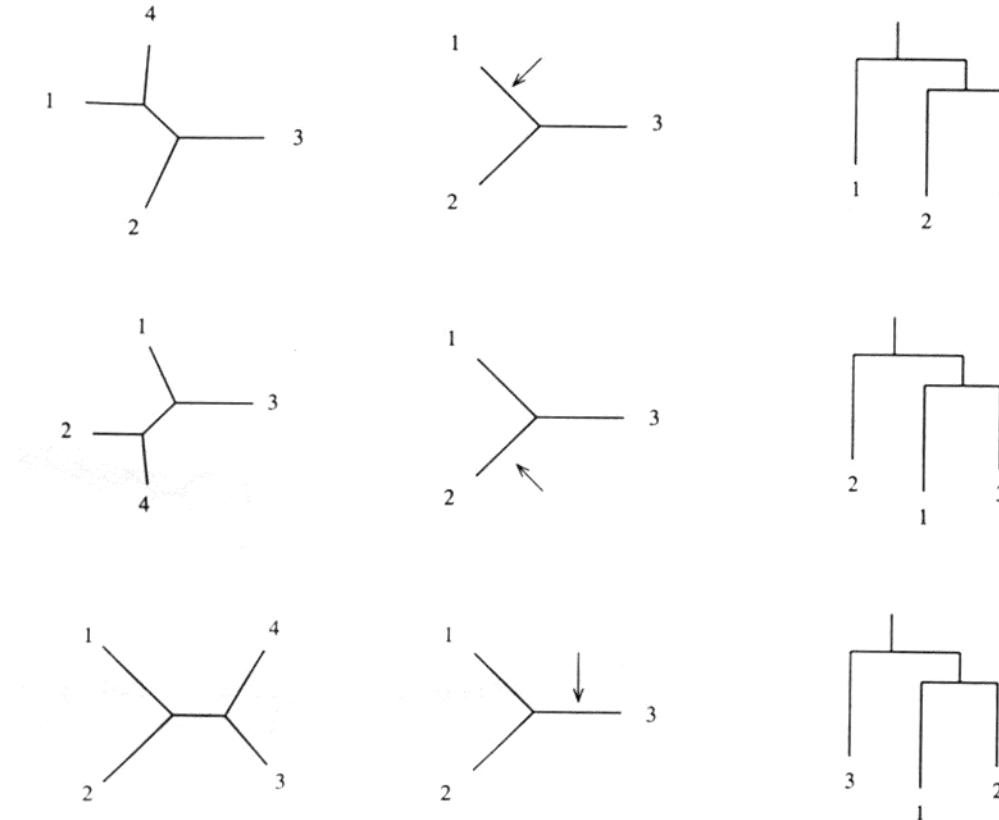
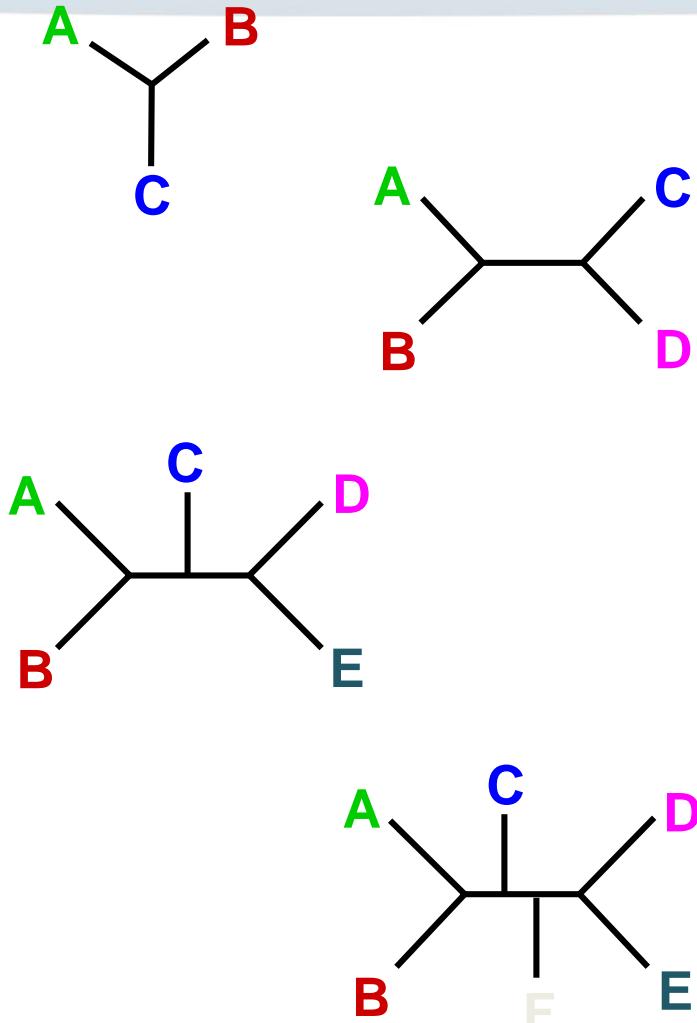


Figure 7.3 The rooted trees (right-hand column) derived from the unrooted tree for three sequences by picking different edges as positions for the root (arrows).

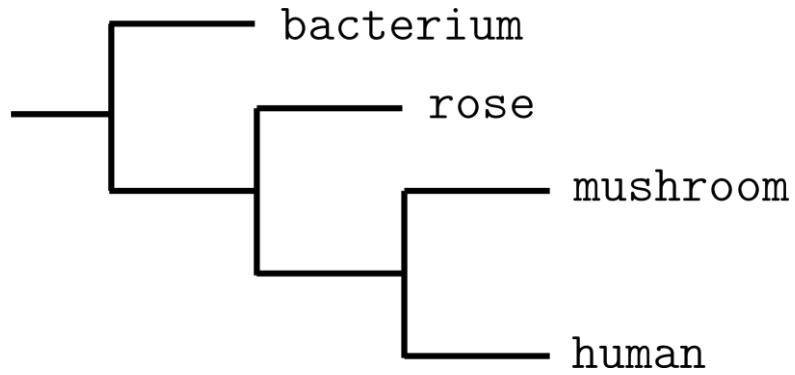
Counting Trees



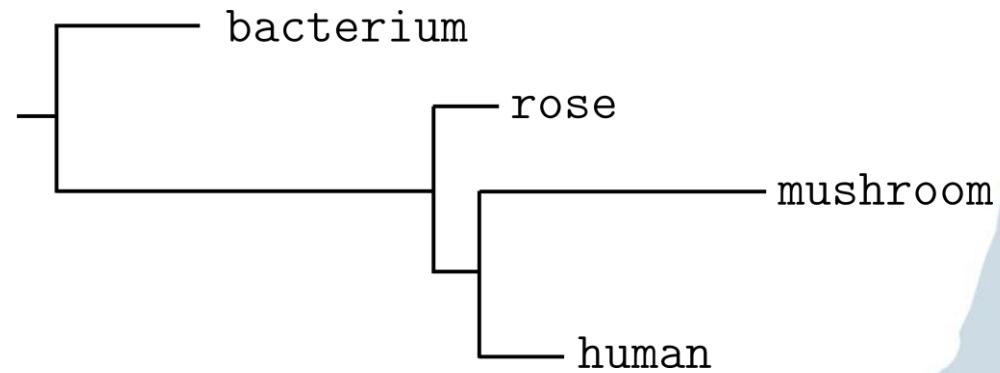
# Taxa (N)	# Unrooted trees
3	1
4	3
5	15
6	105
7	945
8	10,935
9	135,135
10	2,027,025
•	•
•	•
•	•
•	•
30	$\approx 3.58 \times 10^{36}$

$(2N - 5)!! = \# \text{ unrooted trees for } N \text{ taxa}$
 $(2N - 3)!! = \# \text{ rooted trees for } N \text{ taxa}$

Cladogram vs Phylogram

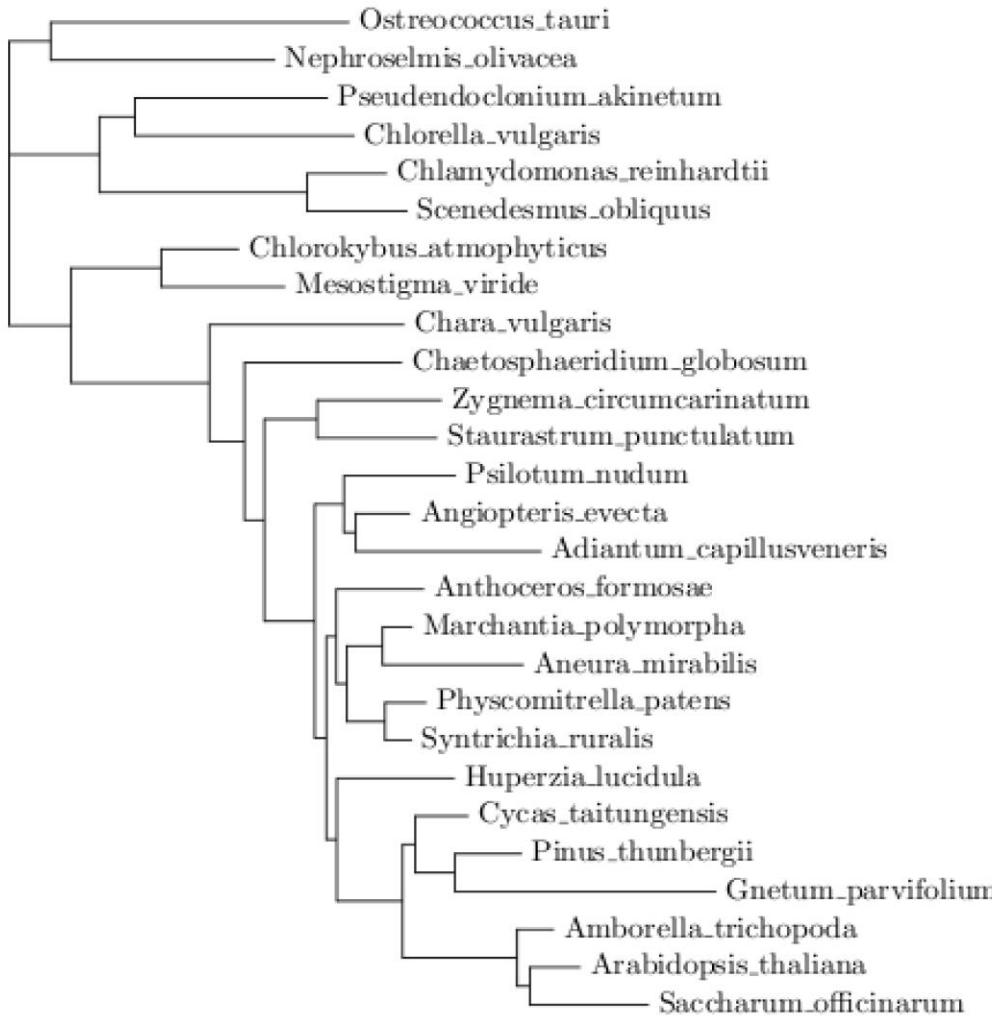


cladogram – arbitrary length branches



phylogram – branch length
proportional
to some measure of genetic distance

Outgroups



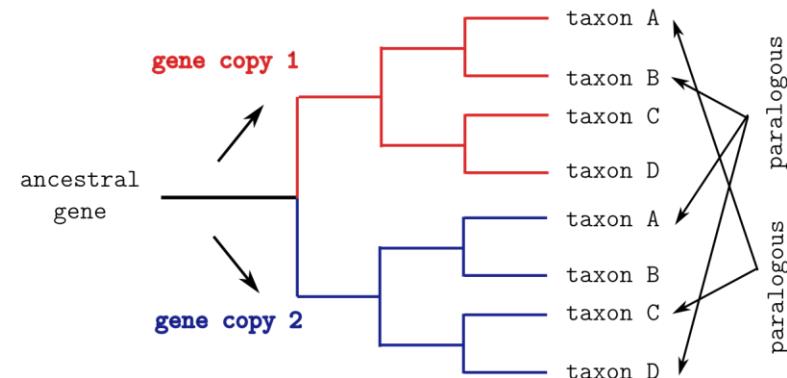
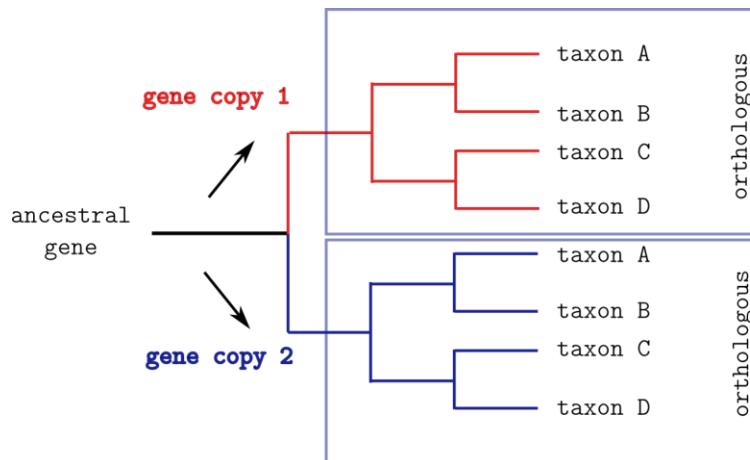
outgroup

ingroup

Orthologs and Paralogs

Orthology – homologous gene sequences

Paralogy – gene sequences separated by a gene duplication event



Rooting

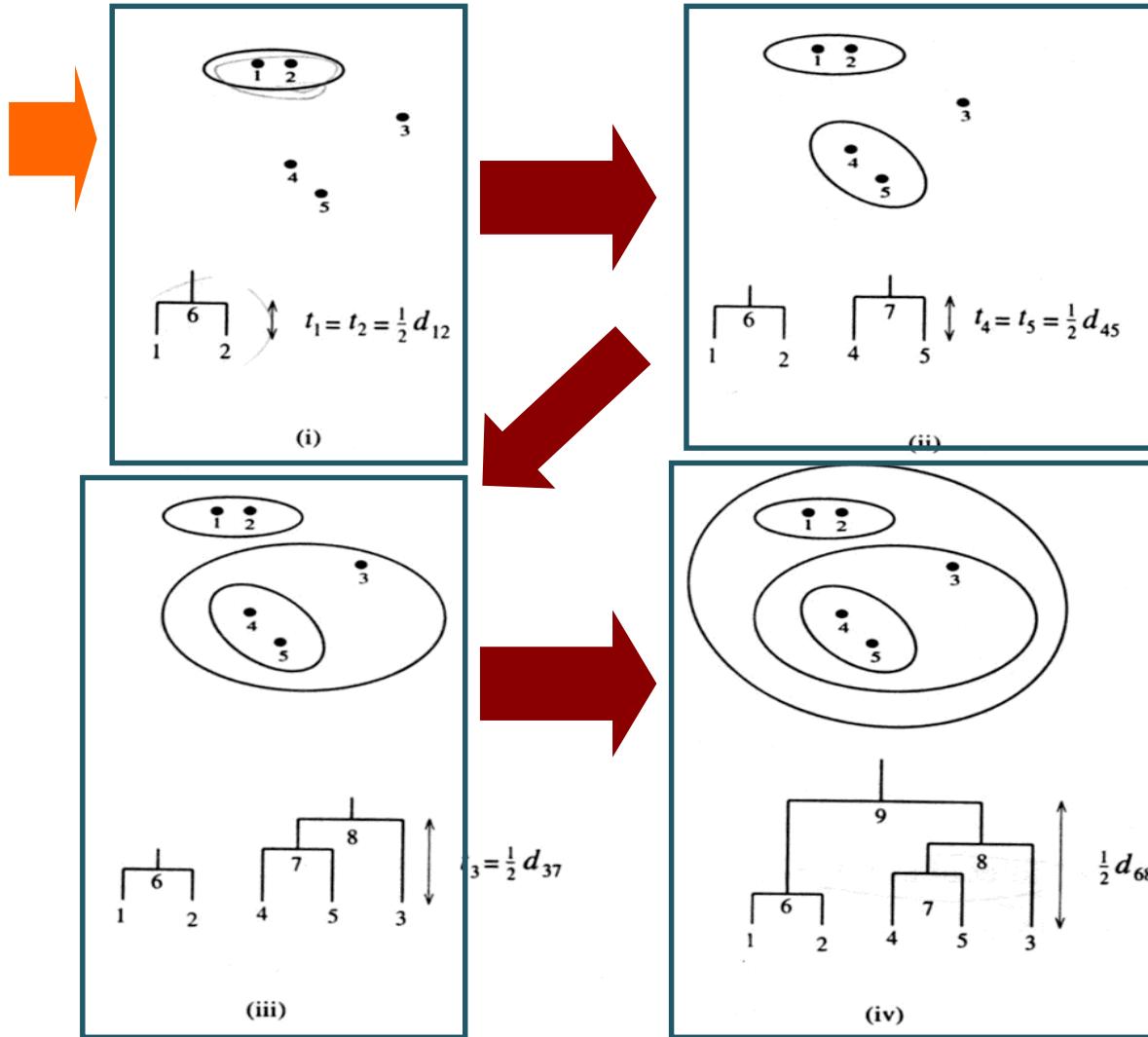
- A phylogenetic analysis most often results in a formally unrooted network
- For phylogenetic analyses can be included an “outgroup” which will be used to root the tree
- The taxa of interested in are called the “ingroup”
- The assumption is that the ingroup taxa are more closely related to each other than any is to the outgroup
- If this assumption is wrong, then the interpretations of the phylogenetic tree will be wrong!

Methods

- Distance-based methods:
- Neighbor-joining
- UPGMA

- Character-based methods:
- Maximum parsimony
- Model-based methods:
 - Maximum likelihood
 - Bayesian inference

UPGMA



UPGMA

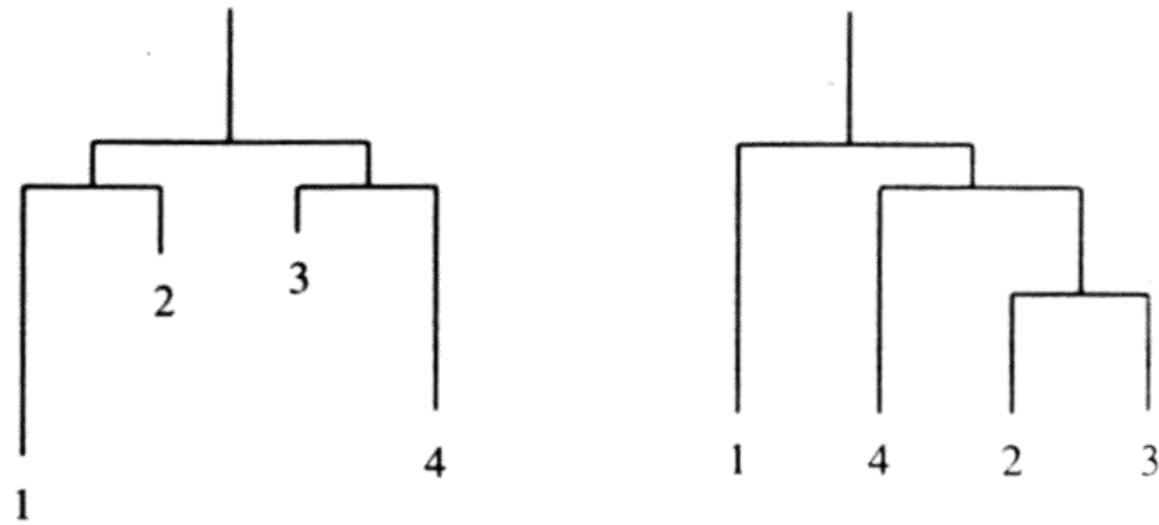
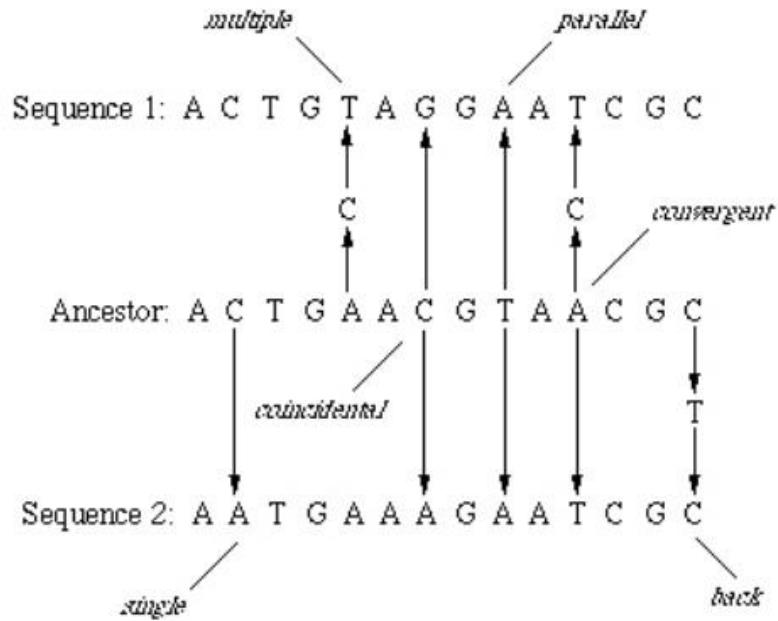


Figure 7.5 A tree (left) that is reconstructed incorrectly by UPGMA (right).

Neighbor Joining

- Very popular method
- Does not make molecular clock assumption : modified distance matrix constructed to adjust for differences in evolution rate of each taxon
- Produces unrooted tree
- Assumes additivity: distance between pairs of leaves = sum of lengths of edges connecting them

Neighbor Joining



Use models of substitution to correct these values

File Formats

- Newick (.nwk)

((species1:BranchLength,species2)Bootstrap,species 3);

- Advantages:
 - easy to perform
 - quick calculation
 - fit for sequences having high similarity scores
- Disadvantages:
 - the sequences are not considered as such (loss of information)
 - all sites are generally equally treated (do not take into account differences of substitution rates)
 - not applicable to distantly divergent sequences.

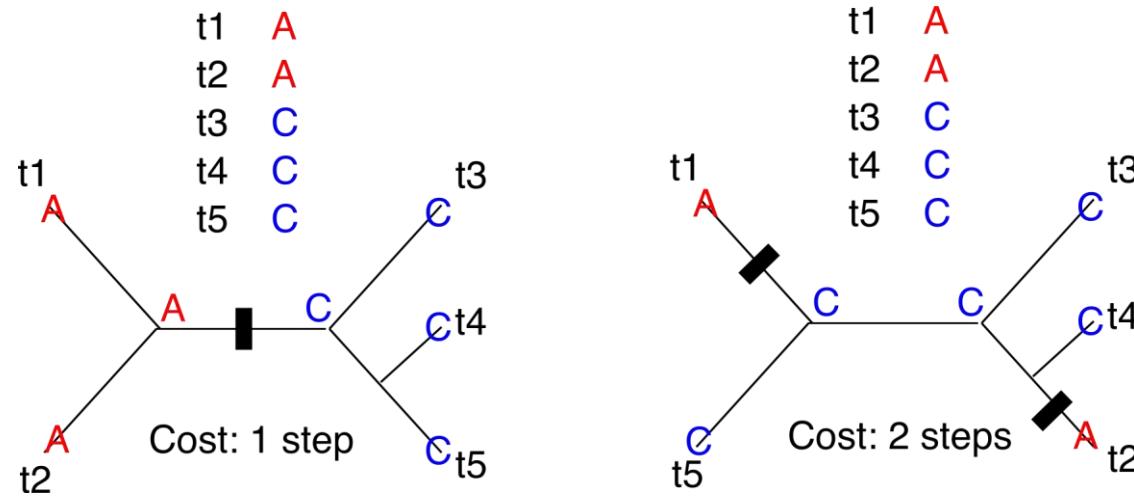
1. Requisitos Seaview e Mega
2. Descarregar sequencias na pasta hands_on_5
3. Alinhar usando Muscle em Aminoácidos
4. Fazer uma NJ de nucleótidos usando os parâmetros por omissão, 1000 bootstrap
5. Repetir o passo fazendo uma árvore de aminoácidos

Characters based tree

Maximum parsimony

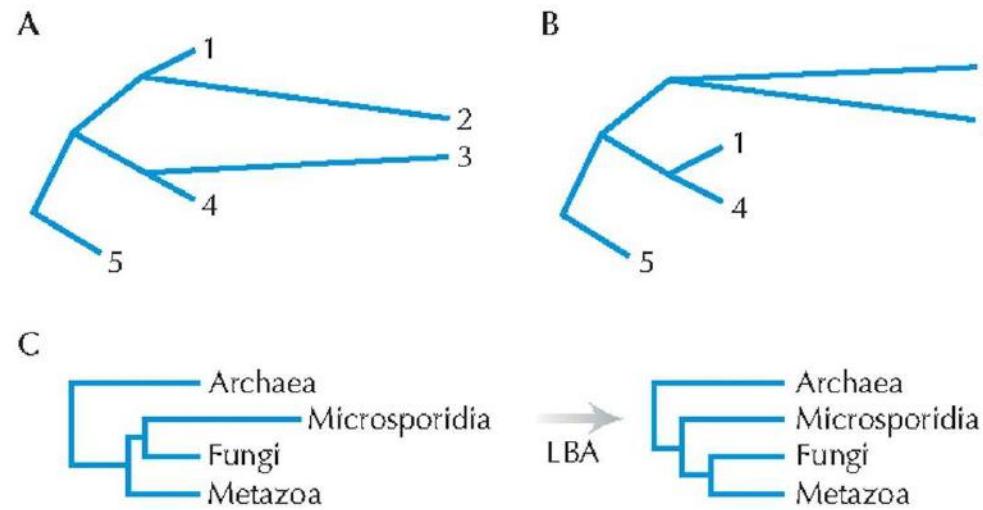
- Object is to **minimise the number of changes** necessary for the evolution of all characters on a tree.
- Character changes are *typically* treated as equally-weighted ie. the "cost" of changing from one state to another is the same between all states, but various weighting schemes can be applied
- Can be used with both morphological and molecular data, morphological characters may be ordered and polarised
- The tree with the **fewest changes/steps** is the MP tree. Might find many most- parsimonious solutions, which are often presented as a 50% majority-rule tree

Maximum parsimony



.Fitch's algorithm can be used to determine the most-parsimonious character reconstruction on any tree, the total score (or length) of a tree is the number of steps (changes) required by the most-parsimonious reconstructions of all characters, and the tree (or trees) with the lowest total score is the MP tree (or MP trees)

Long Branch Attraction



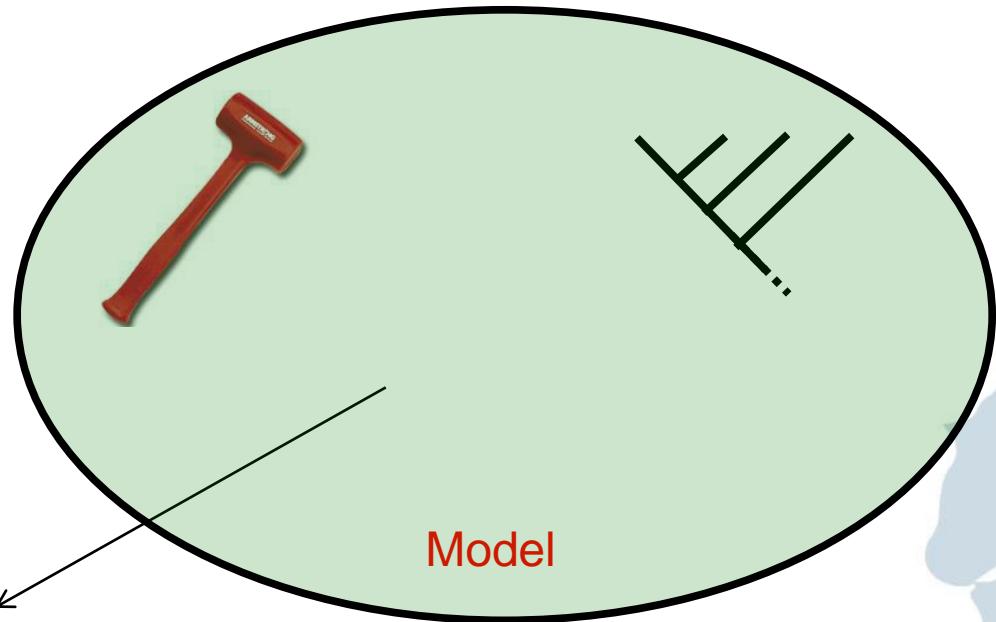
Maximum likelihood and Bayesian inference

- .Both use explicit models of character change that are evaluated on a tree using the likelihood function $\propto \mathcal{P}(\mathcal{D} | \mathcal{H})$
- .They differ in their use of statistical paradigms

Model as mechanism of change and tree



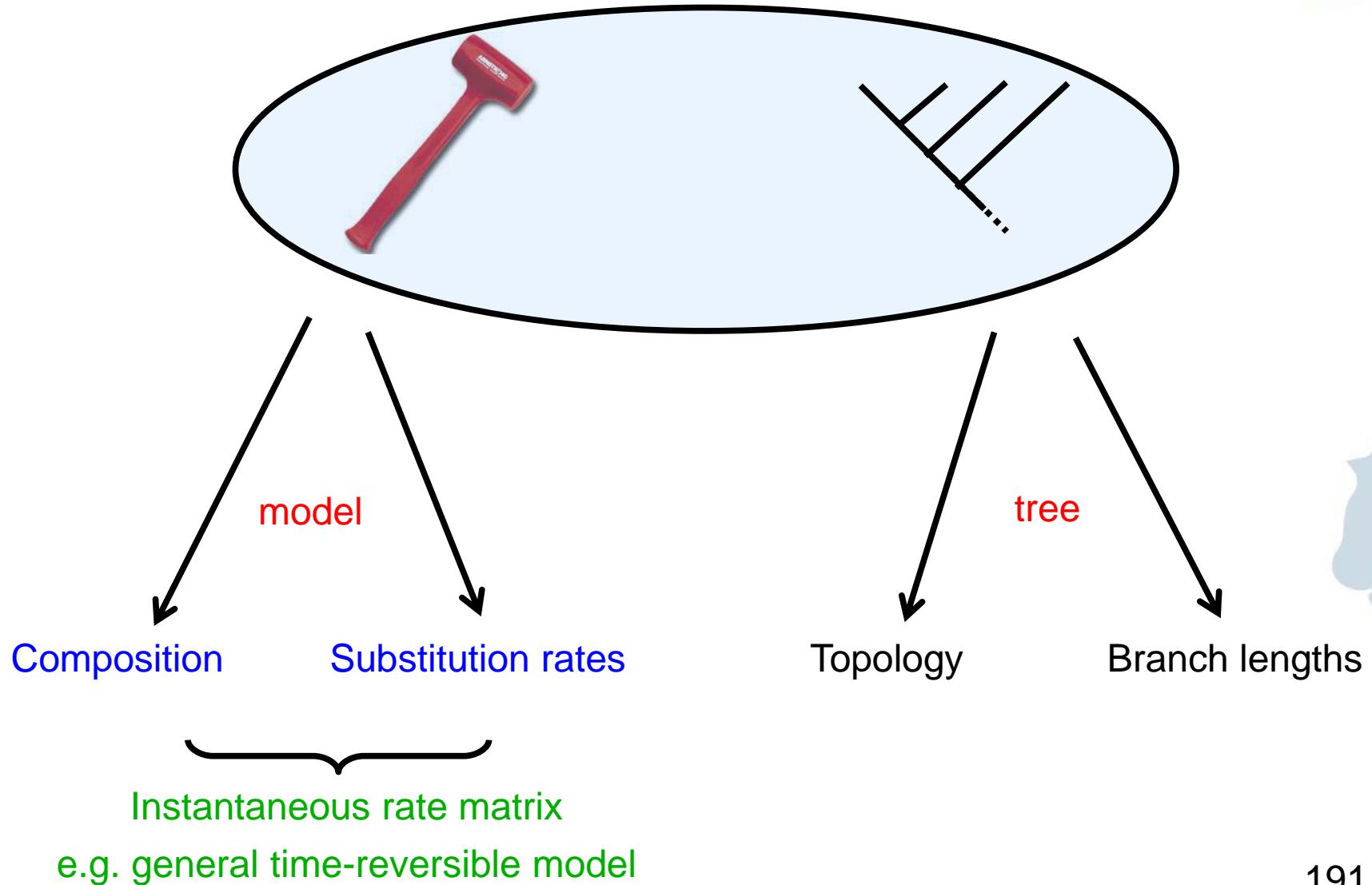
Data



$$\text{likelihood} \propto \text{Prob}(\mathcal{D} | \mathcal{H})$$

The likelihood is proportional to the probability of data given the hypothesis
(a model of character change plus tree topology)

Model as mechanism of change and tree



Models

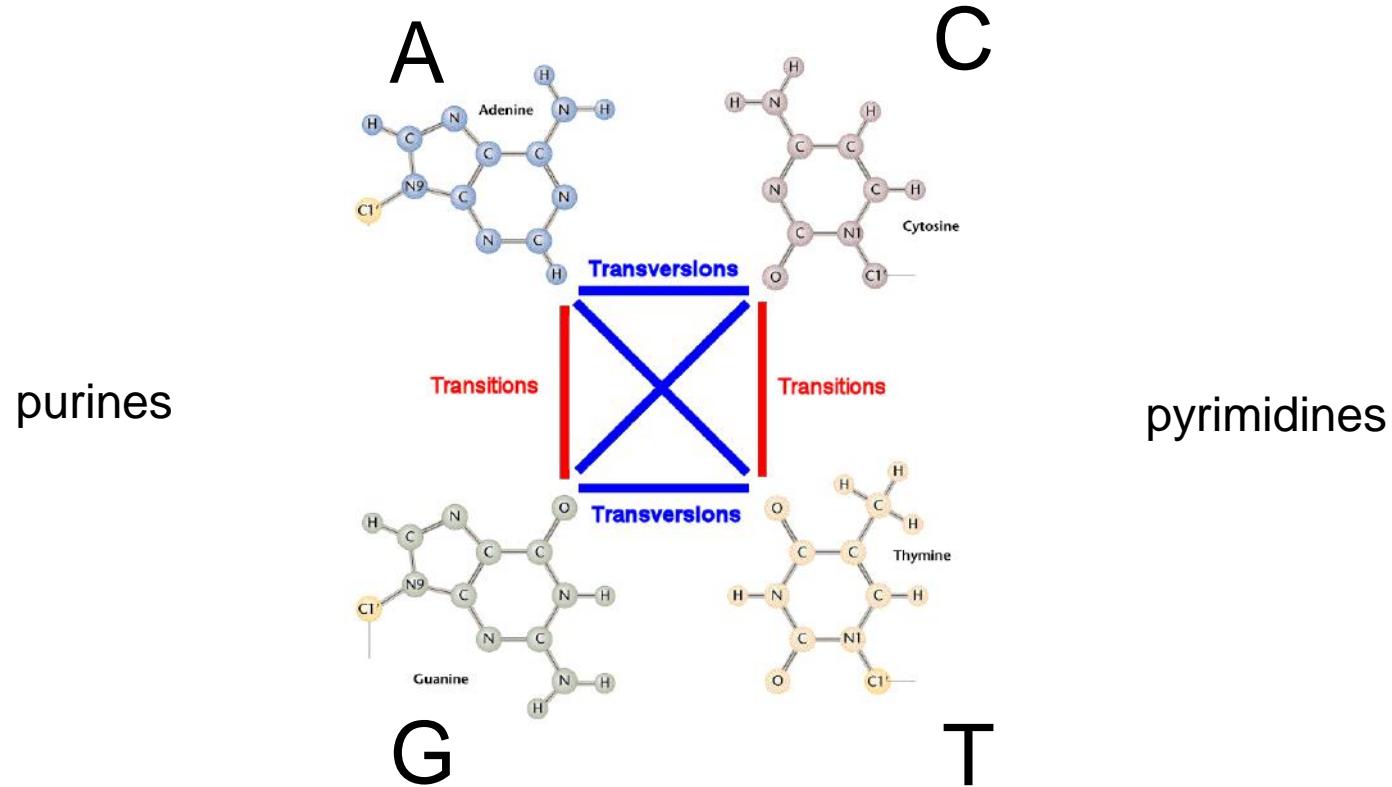
- 1) the base composition:
- π_a , π_c , π_g , & π_t
- The base composition frequency parameters remain constant over time (i.e. they are at equilibrium)
- They express the rate at which changes **to** each base occur
- Hence, the rate of change to a rare base would be low, whereas the change to a common base would be high
 - » could be equal: 0.25, 0.25, 0.25, 0.25
 - » or not: 0.3, 0.4, 0.2, 0.1
 - » perhaps values are estimated from the data

Models

2) Substitution rates:

$$\mathcal{R} = \begin{pmatrix} A & C & G & T \\ - & r_{A \rightarrow C} & r_{A \rightarrow G} & r_{A \rightarrow T} \\ r_{C \rightarrow A} & - & r_{C \rightarrow G} & r_{C \rightarrow T} \\ r_{G \rightarrow A} & r_{G \rightarrow C} & - & r_{G \rightarrow T} \\ r_{T \rightarrow A} & r_{T \rightarrow C} & r_{T \rightarrow G} & - \end{pmatrix} \begin{matrix} A \\ C \\ G \\ T \end{matrix}$$

Transitions and transversions



Model GTR

$$Q = \begin{pmatrix} - & \mu r_{A \rightarrow C} \pi_C & \mu r_{A \rightarrow G} \pi_G & \mu r_{A \rightarrow T} \pi_T \\ \mu r_{C \rightarrow A} \pi_A & - & \mu r_{C \rightarrow G} \pi_G & \mu r_{C \rightarrow T} \pi_T \\ \mu r_{G \rightarrow A} \pi_A & \mu r_{G \rightarrow C} \pi_C & - & \mu r_{G \rightarrow T} \pi_T \\ \mu r_{T \rightarrow A} \pi_A & \mu r_{T \rightarrow C} \pi_C & \mu r_{T \rightarrow G} \pi_G & - \end{pmatrix}$$

$$GTR = \begin{matrix} & \textcolor{red}{A} & \textcolor{red}{C} & \textcolor{red}{G} & \textcolor{red}{T} \\ \textcolor{red}{A} & - & \mu r_i \pi_C & \mu r_j \pi_G & \mu r_k \pi_T \\ \textcolor{red}{C} & \mu r_i \pi_A & - & \mu r_l \pi_G & \mu r_m \pi_T \\ \textcolor{red}{G} & \mu r_j \pi_A & \mu r_l \pi_C & - & \mu r_n \pi_T \\ \textcolor{red}{T} & \mu r_k \pi_A & \mu r_m \pi_C & \mu r_n \pi_G & - \end{matrix}$$

GTR derived

Model derived from the *GTR* model – few have been implemented in phylogenetics

- **GTR** – unequal base frequencies and 6 substitution types
- **SYM** – equal base frequencies and 6 substitution types
- **HKY85** – unequal base frequencies and 2 substitution types (**transitions** and **transversion**)
- **F81** – unequal base frequencies and single substitution type
- **JC** – equal base frequencies and single substitution type

Model selection

2) The **Akaike Information Criterion**: $AIC_i = -2\log L_i + 2p_i$

where: i is the hypothesis (model + tree), and
 p is the number of free parameters

- Does not require models to be nested
- Calculate AIC for each model
- Choose model with lowest AIC score
- To be preferred over the LRT

Model selection

1) The **Likelihood Ratio Test** (LRT):

where: L_0 is a restricted (simpler) version of L_1

e.g.:

Null model = HKY+G

$-\ln L_0 = 7918.9556$

Alternative model = GTR+G

$-\ln L_1 = 7907.3330$

$2(\ln L_1 - \ln L_0) = 23.2451$

$df = 4$

P-value = 0.000113

Tools

- **Modeltest** conducts the LRT/AIC (and others) among a set of nested models
 - 14 substitution matrices with and without a *pinvar* and *gdasrv*
- Uses a crude distance tree to calculate the likelihoods of the models
- **MrModeltest** calculates similar for 24 models (those used by MrBayes)
- For amino acid models use **ProtTest** or **ModelGenerator**

Models parameters

- It is commonly recognised that not all sites evolve at the same rate – some may be constrained by selection. This can be incorporated into the model:
 1. *Site-specific rate categories* - defined *a priori*, e.g. first, second, third codon positions of a protein-coding gene.
 2. *Proportion of invariant sites* (pinvar) - assumes some proportion of sites is incapable of changing and all other sites vary at the same rate.
 3. *Gamma distributed among site rate variation* (gdasrv) - uses a number of discrete categories of rates that partitions a gamma distribution - the shape of the distribution is described by the parameter α

Amino acid substitution models

- **Poisson model** – (equiv. to JC) equal substitution rates and frequencies
- **Proportional model** – as Poisson but with unequal (empirically observed) frequencies

Empirically observed transition matrices:

- **Dayhoff** - derived from Dayhoff, et al.'s (1978) empirical substitution matrix
- **JTT** - Jones, Taylor, Thornton
- **WAG** - Whelan and Goldman

Bootstrap

- The bootstrap is a statistical method that is designed to test the reliability of the result by using **pseudo-replicates drawn from the original data**
- Draw characters/sites from the original data, with replacement, from the original data set to make a new one the same size. Repeat the phylogenetic analysis on this bootstrap replicate and repeat the process many times (100-1000).
- Interpretation of the bootstrap is difficult. It is known to be biased, but for a particular support values it not known where it is biased up or down. Its usually reckoned that 70% is statistically significant.

Bootstrap

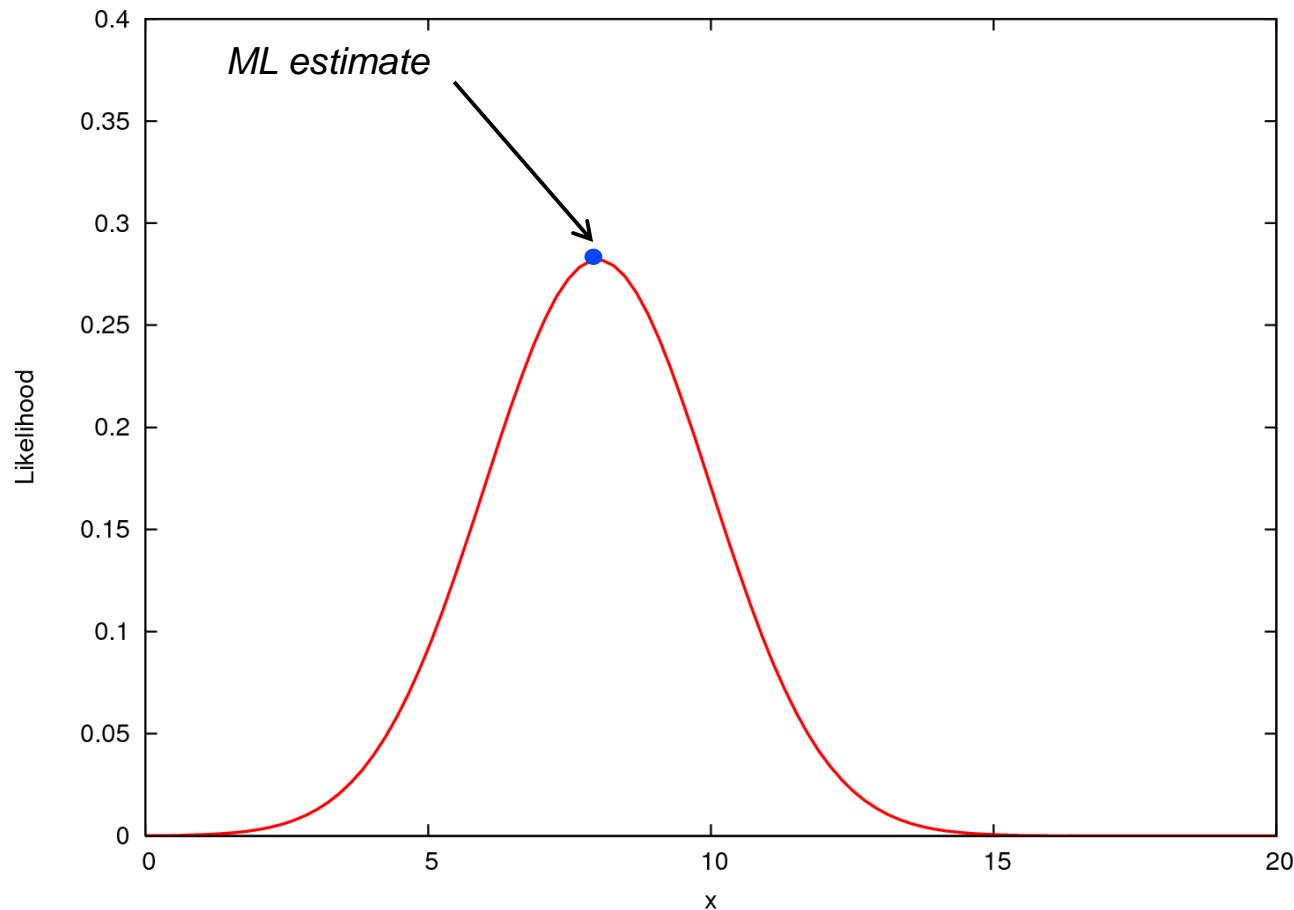
FGFR1_10 -----
 FGFR1_12 -----
 FGFR1_15 MWSWKCLLFWAFLVLTATLCTARPSPLPBC-----
 FGFR1_6 MWSWKCLLFWAFLVLTATLCTARPSPLPBC-----
 FGFR1_8 MWSWKCLLFWAFLVLTATLCTARPSPLPBC-----
 FGFR1_14 MWSWKCLLFWAFLVLTATLCTARPSPLPBCAQI PWGAPVEVESFLVHPGDLQLRCRIRDDVQSINWL RDGVQLAESNRTRITGEEVEVQDSVPADEGLYACVTSSPGSDDTTFSVNVSDALPSSEDDDDDDSSSEEKEIDNTKPNRMP-----
 FGFR1_4 MWSWKCLLFWAFLVLTATLCTARPSPLPBCAQI PWGAPVEVESFLVHPGDLQLRCRIRDDVQSINWL RDGVQLAESNRTRITGEEVEVQDSVPADEGLYACVTSSPGSDDTTFSVNVSDALPSSEDDDDDDSSSEEKEIDNTKPNRMP-----
 FGFR1_1 MWSWKCLLFWAFLVLTATLCTARPSPLPBCAQI PWGAPVEVESFLVHPGDLQLRCRIRDDVQSINWL RDGVQLAESNRTRITGEEVEVQDSVPADEGLYACVTSSPGSDDTTFSVNVSDALPSSEDDDDDDSSSEEKEIDNTKPNRMP-----
 FGFR1_11 -----
 FGFR1_13 -----
 FGFR1_7 MWSWKCLLFWAFLVLTATLCTARPSPLPBC-----
 FGFR1_9 MWSWKCLLFWAFLVLTATLCTARPSPLPBC-----
 FGFR1_17 MWSWKCLLFWAFLVLTATLCTARPSPLPBC-----
 FGFR1_18 MWSWKCLLFWAFLVLTATLCTARPSPLPBC-----
 FGFR1_2 MWSWKCLLFWAFLVLTATLCTARPSPLPBCAQI PWGAPVEVESFLVHPGDLQLRCRIRDDVQSINWL RDGVQLAESNRTRITGEEVEVQDSVPADEGLYACVTSSPGSDDTTFSVNVSDALPSSEDDDDDDSSSEEKEIDNTKPNRMP-----
 FGFR1_5 MWSWKCLLFWAFLVLTATLCTARPSPLPBCAQI PWGAPVEVESFLVHPGDLQLRCRIRDDVQSINWL RDGVQLAESNRTRITGEEVEVQDSVPADEGLYACVTSSPGSDDTTFSVNVSDALPSSEDDDDDDSSSEEKEIDNTKPNRMP-----
 FGFR1_3 MWSWKCLLFWAFLVLTATLCTARPSPLPBCAQI PWGAPVEVESFLVHPGDLQLRCRIRDDVQSINWL RDGVQLAESNRTRITGEEVEVQDSVPADEGLYACVTSSPGSDDTTFSVNVSDALPSSEDDDDDDSSSEEKEIDNTKPNRMP-----
 FGFR1_16 MWSWKCLLFWAFLVLTATLCTARPSPLPBCAQI PWGAPVEVESFLVHPGDLQLRCRIRDDVQSINWL RDGVQLAESNRTRITGEEVEVQDSVPADEGLYACVTSSPGSDDTTFSVNVSDALPSSEDDDDDDSSSEEKEIDNTKPNRMP-----
 ruler 1.....10.....20.....30.....40.....50.....60.....70.....80.....90.....100.....110.....120.....130.....140.....150

FGFR1_10 -----MEKKLHAVPAAKTVIFKCPSSGTIPNPTLRLWLNKGKEFKPDHRIGGYKVRYATWSIIMDSVVPSDKGNYTCIVENEYGSINHTYQLDVVERSPHRPILOAGLPANKTVALGSNVFMCVKVYSPOPHQIWLKHIEVNGSKI
 FGFR1_12 -----MEKKLHAVPAAKTVIFKCPSSGTIPNPTLRLWLNKGKEFKPDHRIGGYKVRYATWSIIMDSVVPSDKGNYTCIVENEYGSINHTYQLDVVERSPHRPILOAGLPANKTVALGSNVFMCVKVYSPOPHQIWLKHIEVNGSKI
 FGFR1_15 VAPYWTSPPEKMEKKLHAVPAAKTVIFKCPSSGTIPNPTLRLWLNKGKEFKPDHRIGGYKVRYATWSIIMDSVVPSDKGNYTCIVENEYGSINHTYQLDVVERSPHRPILOAGLPANKTVALGSNVFMCVKVYSPOPHQIWLKHIEVNGSKI
 FGFR1_6 VAPYWTSPPEKMEKKLHAVPAAKTVIFKCPSSGTIPNPTLRLWLNKGKEFKPDHRIGGYKVRYATWSIIMDSVVPSDKGNYTCIVENEYGSINHTYQLDVVERSPHRPILOAGLPANKTVALGSNVFMCVKVYSPOPHQIWLKHIEVNGSKI
 FGFR1_8 VAPYWTSPPEKMEKKLHAVPAAKTVIFKCPSSGTIPNPTLRLWLNKGKEFKPDHRIGGYKVRYATWSIIMDSVVPSDKGNYTCIVENEYGSINHTYQLDVVERSPHRPILOAGLPANKTVALGSNVFMCVKVYSPOPHQIWLKHIEVNGSKI
 FGFR1_14 VAPYWTSPPEKMEKKLHAVPAAKTVIFKCPSSGTIPNPTLRLWLNKGKEFKPDHRIGGYKVRYATWSIIMDSVVPSDKGNYTCIVENEYGSINHTYQLDVVERSPHRPILOAGLPANKTVALGSNVFMCVKVYSPOPHQIWLKHIEVNGSKI
 FGFR1_4 VAPYWTSPPEKMEKKLHAVPAAKTVIFKCPSSGTIPNPTLRLWLNKGKEFKPDHRIGGYKVRYATWSIIMDSVVPSDKGNYTCIVENEYGSINHTYQLDVVERSPHRPILOAGLPANKTVALGSNVFMCVKVYSPOPHQIWLKHIEVNGSKI
 FGFR1_1 VAPYWTSPPEKMEKKLHAVPAAKTVIFKCPSSGTIPNPTLRLWLNKGKEFKPDHRIGGYKVRYATWSIIMDSVVPSDKGNYTCIVENEYGSINHTYQLDVVERSPHRPILOAGLPANKTVALGSNVFMCVKVYSPOPHQIWLKHIEVNGSKI
 FGFR1_11 -----MEKKLHAVPAAKTVIFKCPSSGTIPNPTLRLWLNKGKEFKPDHRIGGYKVRYATWSIIMDSVVPSDKGNYTCIVENEYGSINHTYQLDVVERSPHRPILOAGLPANKTVALGSNVFMCVKVYSPOPHQIWLKHIEVNGSKI
 FGFR1_13 -----MEKKLHAVPAAKTVIFKCPSSGTIPNPTLRLWLNKGKEFKPDHRIGGYKVRYATWSIIMDSVVPSDKGNYTCIVENEYGSINHTYQLDVVERSPHRPILOAGLPANKTVALGSNVFMCVKVYSPOPHQIWLKHIEVNGSKI
 FGFR1_7 VAPYWTSPPEKMEKKLHAVPAAKTVIFKCPSSGTIPNPTLRLWLNKGKEFKPDHRIGGYKVRYATWSIIMDSVVPSDKGNYTCIVENEYGSINHTYQLDVVERSPHRPILOAGLPANKTVALGSNVFMCVKVYSPOPHQIWLKHIEVNGSKI
 FGFR1_9 VAPYWTSPPEKMEKKLHAVPAAKTVIFKCPSSGTIPNPTLRLWLNKGKEFKPDHRIGGYKVRYATWSIIMDSVVPSDKGNYTCIVENEYGSINHTYQLDVVERSPHRPILOAGLPANKTVALGSNVFMCVKVYSPOPHQIWLKHIEVNGSKI
 FGFR1_17 VAPYWTSPPEKMEKKLHAVPAAKTVIFKCPSSGTIPNPTLRLWLNKGKEFKPDHRIGGYKVRYATWSIIMDSVVPSDKGNYTCIVENEYGSINHTYQLDVVERSPHRPILOAGLPANKTVALGSNVFMCVKVYSPOPHQIWLKHIEVNGSKI
 FGFR1_18 VAPYWTSPPEKMEKKLHAVPAAKTVIFKCPSSGTIPNPTLRLWLNKGKEFKPDHRIGGYKVRYATWSIIMDSVVPSDKGNYTCIVENEYGSINHTYQLDVVERSPHRPILOAGLPANKTVALGSNVFMCVKVYSPOPHQIWLKHIEVNGSKI
 FGFR1_2 VAPYWTSPPEKMEKKLHAVPAAKTVIFKCPSSGTIPNPTLRLWLNKGKEFKPDHRIGGYKVRYATWSIIMDSVVPSDKGNYTCIVENEYGSINHTYQLDVVERSPHRPILOAGLPANKTVALGSNVFMCVKVYSPOPHQIWLKHIEVNGSKI
 FGFR1_5 VAPYWTSPPEKMEKKLHAVPAAKTVIFKCPSSGTIPNPTLRLWLNKGKEFKPDHRIGGYKVRYATWSIIMDSVVPSDKGNYTCIVENEYGSINHTYQLDVVERSPHRPILOAGLPANKTVALGSNVFMCVKVYSPOPHQIWLKHIEVNGSKI
 FGFR1_3 -----CPDLEBAKSCSASFHSI
 FGFR1_16 -----PWGAPVEVBSELVHPGDLQLRCRIRDDVQSINWL RDGVQLAESNRTRITGEEVEVQDSVPADSGLYACVTSSPGSDDTTFSVNVS-----ACPDLEBAKWCASFHSI
 ruler160.....170.....180.....190.....200.....210.....220.....230.....240.....250.....260.....270.....280.....290.....300

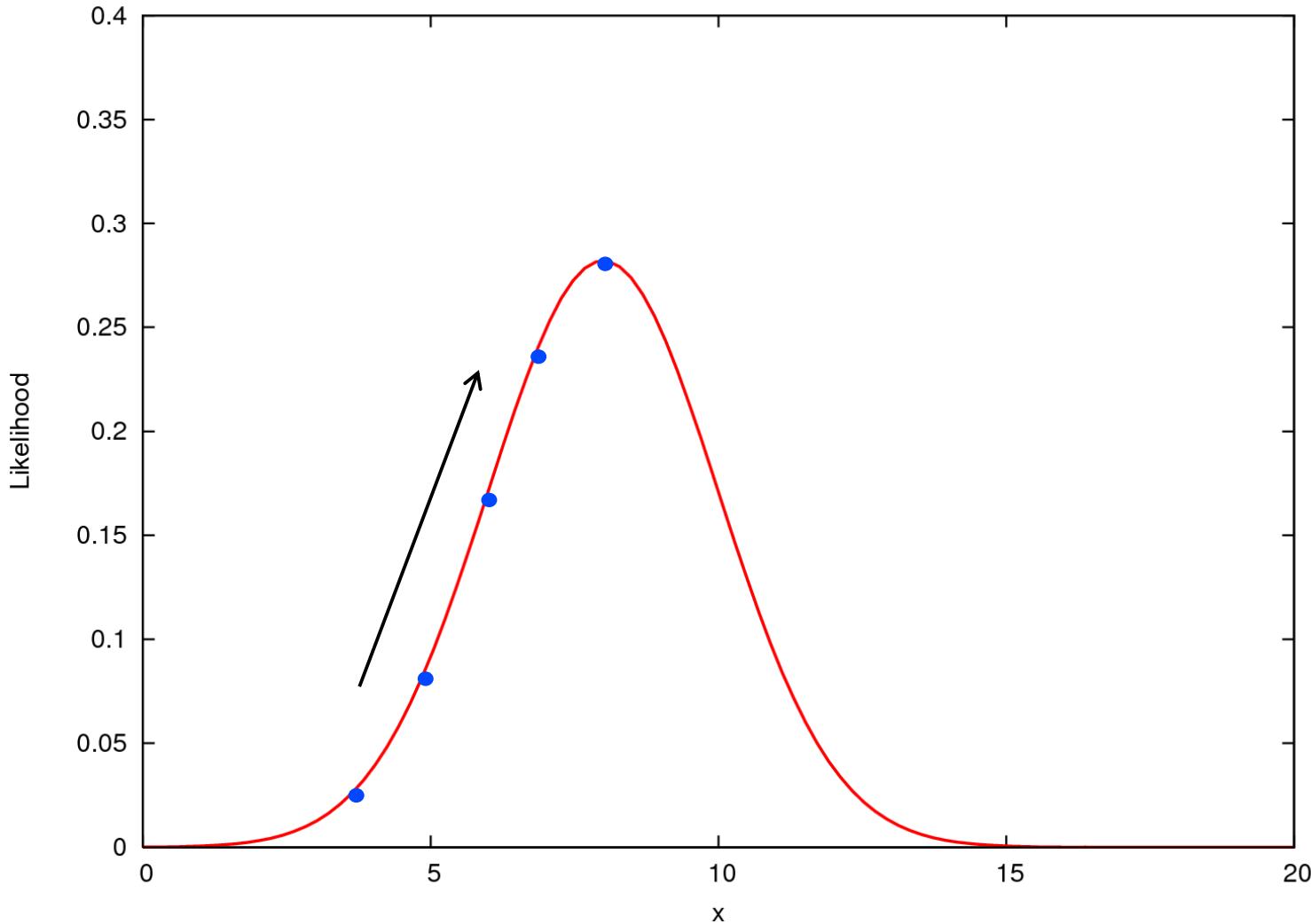
Hands On 6

1. Requisitos Seaview+jModelTest e/ou Mega
2. Alinhar usando MUSCLE
3. Determinar modelo evolucionário
4. Fazer uma ML usando 100 de bootstrap e o modelo determinado no passo 3

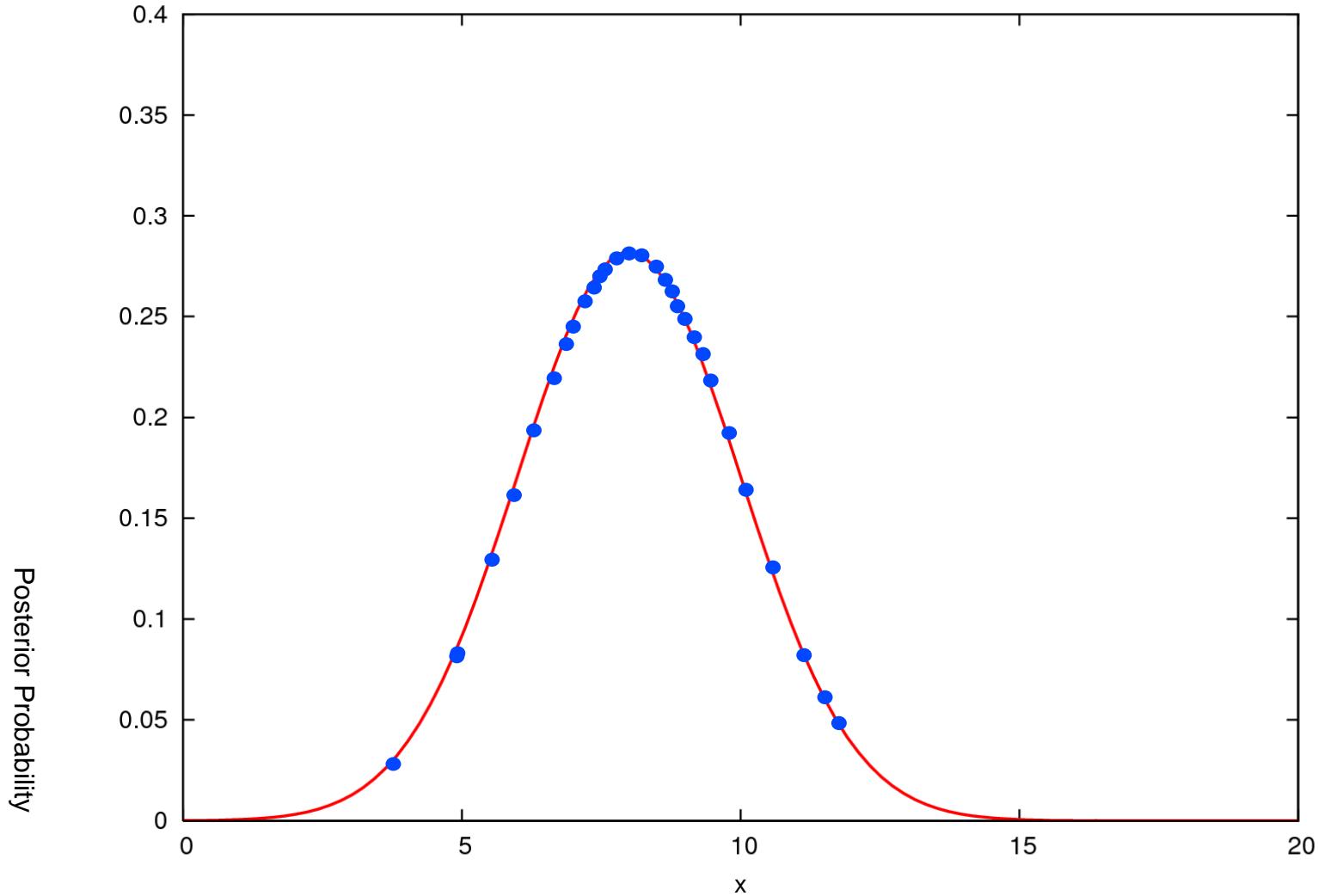
ML and PP density



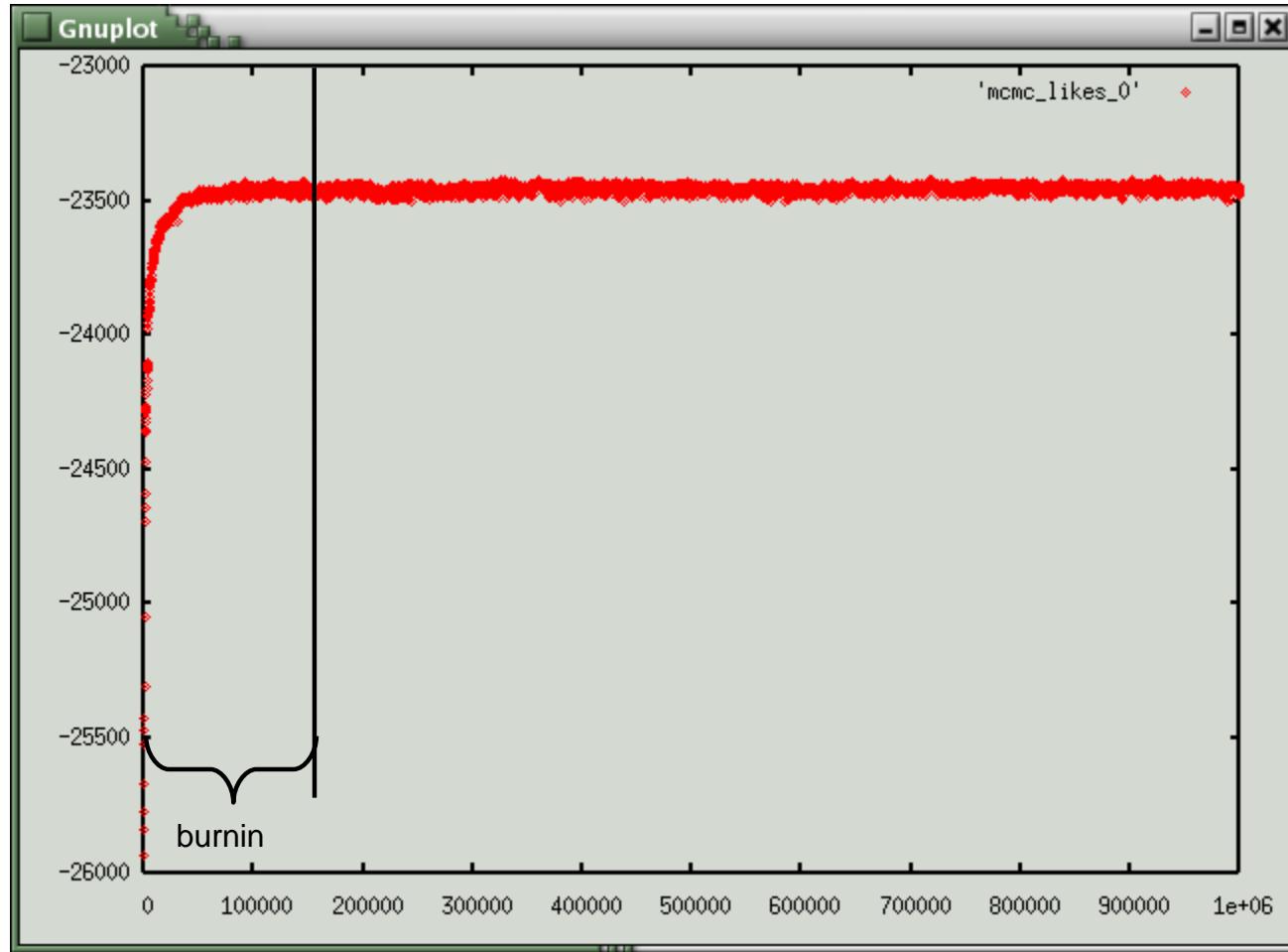
ML and PP density



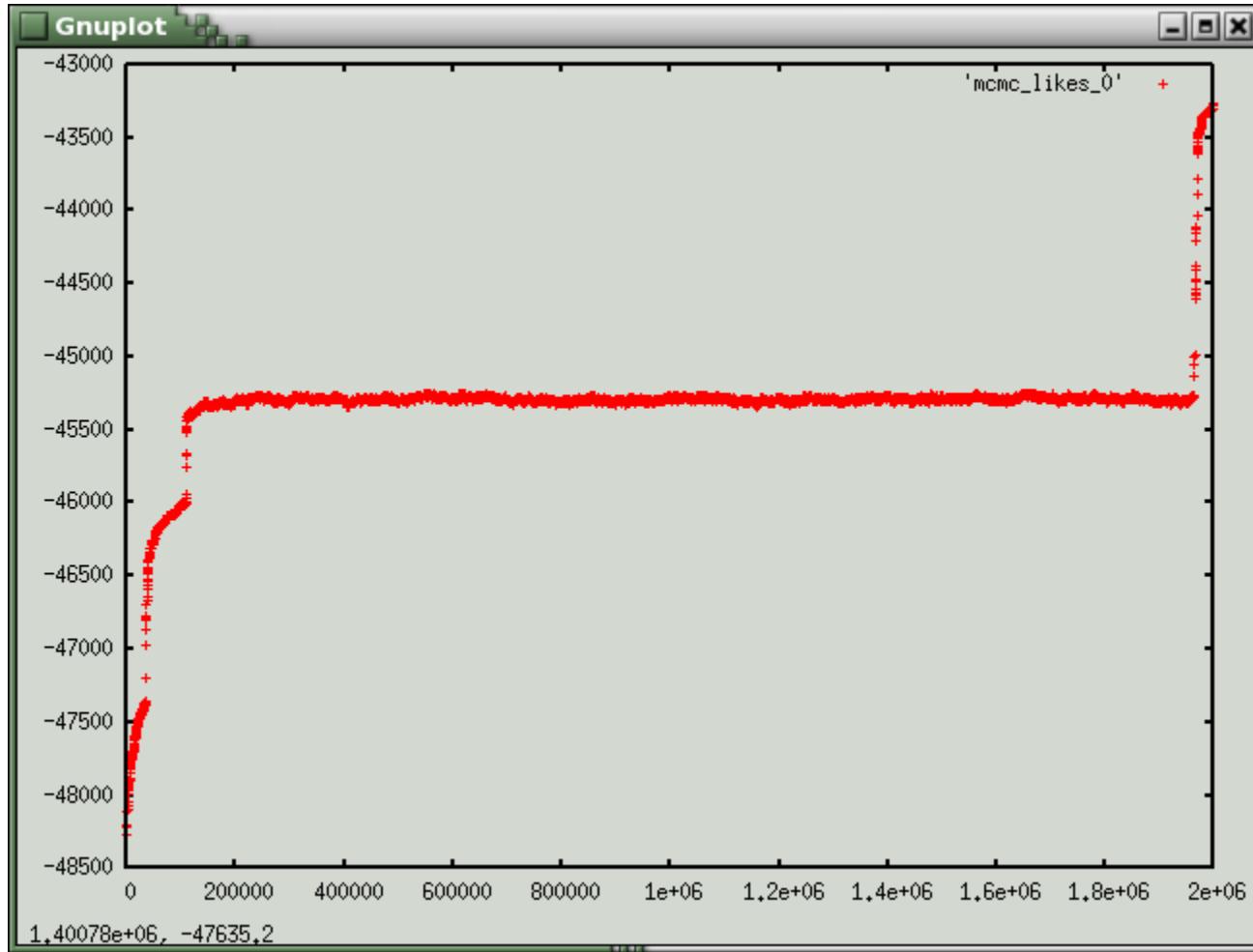
ML and PP density



Likelihood plot



Likelihood plot



Mcmc output

gens	Composition				Substitution Rate					gdasrv	
	A	C	G	T	A-C	A-G	A-T	C-G	C-T	G-T	
→ 100	0.22515466	0.24584661	0.30733981	0.22165891	0.11970749	0.23917488	0.08580185	0.14260800	0.32394051	0.08876728	0.874116
200	0.22581011	0.23935950	0.28160072	0.25322967	0.11361513	0.22751384	0.10471647	0.16653072	0.26859341	0.11903044	1.930917
300	0.23172551	0.27706732	0.27019486	0.22101230	0.12093299	0.20611349	0.08865832	0.15770240	0.28720945	0.13938334	4.380087
400	0.22999880	0.26063025	0.28697633	0.22239461	0.09333566	0.24749557	0.12001019	0.14071384	0.29772513	0.10071961	10.893358
500	0.24616074	0.25080719	0.27680877	0.22622330	0.10333398	0.20508574	0.10527309	0.16562894	0.31464847	0.10602977	14.875529
600	0.21219225	0.28281963	0.28884274	0.21614538	0.11325671	0.25977835	0.12678584	0.11935713	0.28575591	0.09506605	8.521777
etc...											
→ tree t_100 = [&U] (((1:0.263151, 2:0.0564195):0.206267, 3:0.223034):0.73243, 4:0.107335, 5:0.0742962); tree t_200 = [&U] (((5:0.115119, 4:0.0936513):1.14124, 2:0.0686334):0.206054, 3:0.338046, 1:0.223061); tree t_300 = [&U] ((2:0.0799976, (4:0.0892171, 5:0.115119):1.10016):0.174534, 3:0.461212, 1:0.236187); tree t_400 = [&U] ((2:0.0800969, (5:0.126667, 4:0.09186):1.39515):0.21926, 3:0.445261, 1:0.335234); tree t_500 = [&U] (((4:0.100586, 5:0.149979):1.92335, 2:0.0800969):0.268639, 1:0.335234, 3:0.556183); tree t_600 = [&U] (((3:0.671303, 1:0.413087):0.347891, 2:0.178729):2.18893, 5:0.0845881, 4:0.143907); etc...											

- Note that at each generation the parameter values are known (i.e. they are the current values of the chain) hence the likelihood is easy and quick to calculate, this makes BA a relatively quick method when compared to ML

Likelihood plot

- Are we there yet?

AWTY is a system for the graphical exploration of MCMC convergence, written by Jim Wilgenbusch, Dan Warren, and David Swofford.

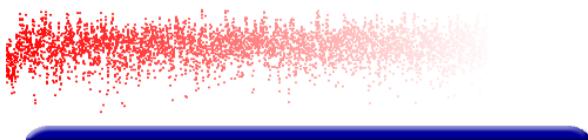
AWTY online

[About AWTY](#)

[Start a new session](#)

[Return to an old session](#)

[Contact us](#)



About AWTY

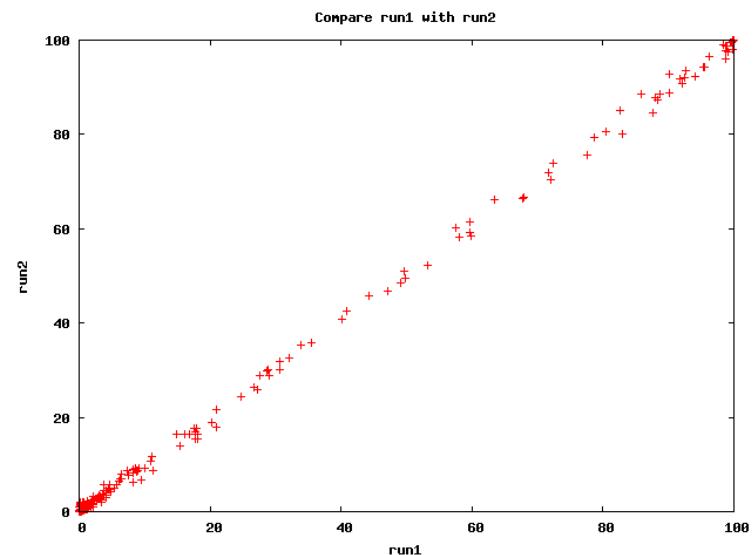
AWTY is a system for graphical exploration of Markov chain Monte Carlo (MCMC) convergence in Bayesian phylogenetic inference. The graphics produced by AWTY are designed to help assess whether an MCMC analysis has run long enough, such that tree topologies are being sampled in proportion to their true posterior probability distribution. In other words, "Are We There Yet?" or AWTY for short. Admittedly, the results generated by AWTY will never be able to answer this question with a definitive yes; however, in some cases results will point confidently to the answer no. See the [AWTY image gallery](#) for some examples.

To produce plots in AWTY a NEXUS or NEWICK formatted tree file representing a set of trees sampled over an MCMC run is required. To date, tree files generated by [MrBayes](#) and [BAMBE](#) have been tested. AWTY provides several graphical formats to display results or results may also be downloaded and analyzed using the plotting package of your choice.

The online version of AWTY is written in [Perl](#) and [PHP](#). Posterior probabilities of splits and topological tree distances are calculated by [PAUP*](#). Graphics are generated by [Gnuplot](#).

Citation:

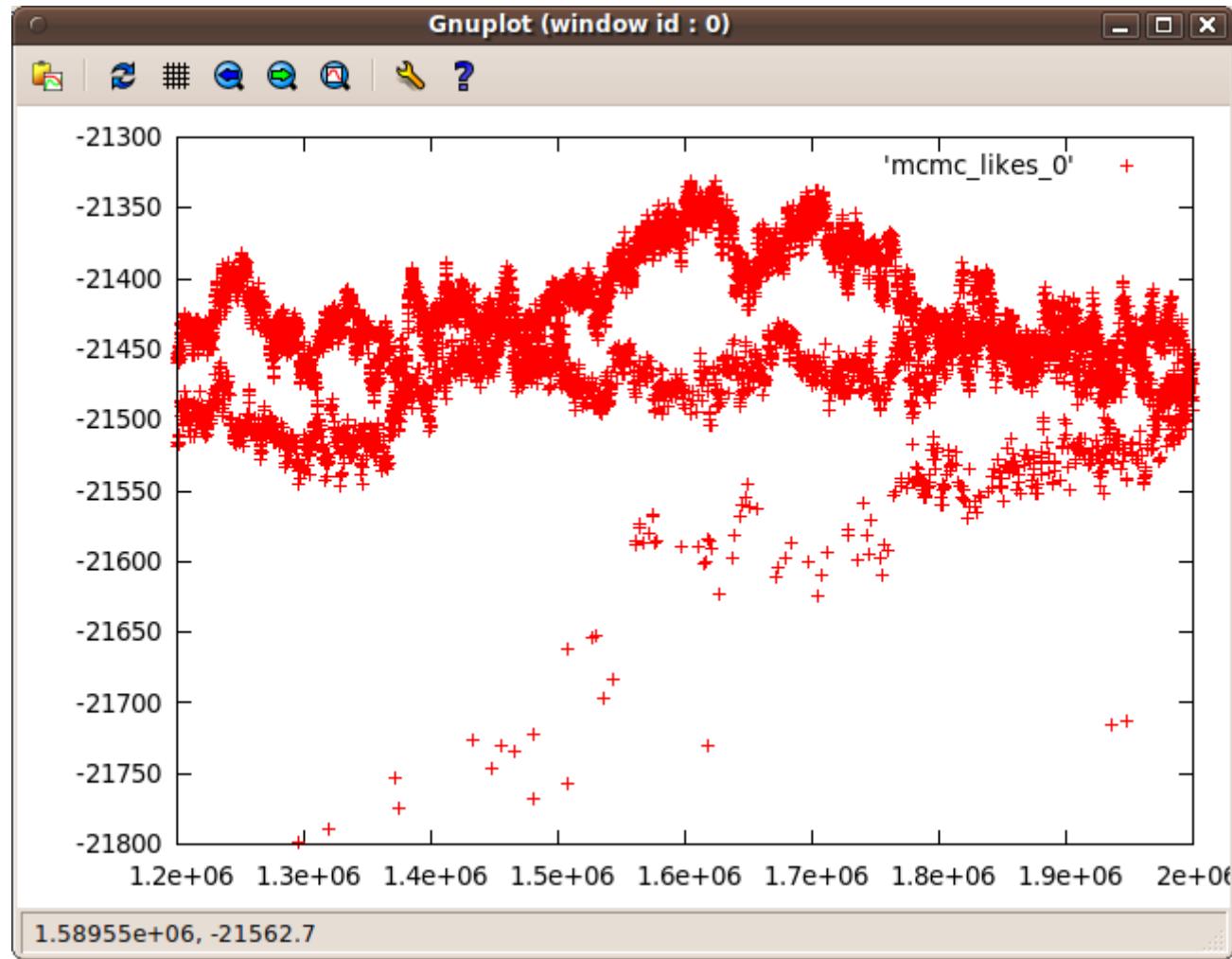
Wilgenbusch J.C., Warren D.L., Swofford D.L. 2004. AWTY: A system for graphical exploration of MCMC convergence in Bayesian phylogenetic inference.
<http://ceb.csit.fsu.edu/awty>.



Convergence

- Run the analysis more than once and check that the separate runs give similar results
- Monitor the average standard deviation of split support between two separate runs (MrBayes does this by default)

Likelihood plot



Alternatives

- <http://www.thines-lab.senckenberg.de/simba/>
- <http://phylemon.bioinfo.cipf.es/>
- http://phylogeny.lirmm.fr/phylo_cgi/index.cgi

Hands On 7

1. Requisitos Mr. Bayes e TreeGraph
2. Abrir executável do Mr.Bayes (32 ou 64 bits)
3. “Execute nomedoficheiro.nex”
4. “help lset”
5. “lset nst=6 rates=invgamma” (modelo GTR+I+G)
6. “help mcmc”
7. “mcmc ngen=500000”
8. “sump” (confirmar parâmetros)
9. “sumt conformat=simple”
10. Executar TreeGraph e abrir ficheiro .con gerado no Mr. Bayes

Hands On 8

1. Requisitos Tree-Puzzle
2. Abrir puzzle-windows-mingw.exe
3. k + enter -> K Tree search procedure? Evaluate user defined trees
4. m + enter até ao modelo GTR
5. w + enter até -> w Model of rate heterogeneity? Mixed (1 invariable + 4 Gamma rates)
6. y + enter
7. my_trees.txt (ficheiros com as árvores)
8. Abrir ficheiro .puzzle
9. No final do ficheiro aparece a comparação entre as 3 árvores “COMPARISON OF USER TREES (NO CLOCK)”

Data Warehouses Again

EMBL-EBI Treefam

Services | Research | Training | About us

Search Examples: BRCA2, ENSP0000428982, or do a sequence search

Home | Search | Browse | Download | Help | Forum

Family: awaiting annotation (TF337278)
Description: awaiting annotation

Summary

- Gene Tree
- Wikipedia
- Sequences
- Downloads

Summary

Family info

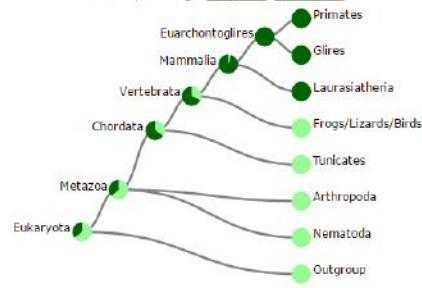
Name:	awaiting annotation
Accession:	TF337278
Description:	awaiting annotation
Taxonomic distribution:	Metazoa

Domain(s) and Function(s)

-  Enamelin (100% of seqs.)
-  HGNC: FVAM

Which species have awaiting annotation?

show percentage: by species/ by sequence



Legend: dark green shows present species/genes. Light green shows missing species/genes.

EMBL-EBI

Services

- By topic
- By name (A-Z)
- Help & Support

Research

- Overview
- Publications
- Research groups
- Postdocs & PhDs

Training

- Overview
- Train at EBI
- Train outside EBI
- Train online
- Contact organisers

Industry

- Overview
- Members Area
- Workshops
- SME Forum
- Contact Industry programme

About us

- Overview
- Leadership
- Funding
- Background
- Collaboration
- Jobs
- People & groups
- News

Data Warehouses Again

e!Ensembl BLAST/BLAT | BioMart | Tools | Downloads | Help & Documentation | Blog | Mirrors

Search: All species for Go

e.g. BRCA2 or rat 5:62797383-63627669 or rs699 or coronary heart disease

Browse a Genome
The Ensembl project produces genome databases for vertebrates and other eukaryotic species, and makes this information freely available online.

Popular genomes

 Human GRCh38.p6	 Human GRCh37
 Mouse GRCm38.p4	 Zebrafish GRCz10

[★ Log in to customize this list](#)

All genomes

[View full list of all Ensembl species](#)

Other species are available in [Ensembl Protocols](#) and [Ensembl Genomes](#)

New! From the 24th March – 5th May you can take part in our [interactive online webinar course](#), and afterwards catch up the recordings

Still using Human GRCh37? [Go to GRCh38](#)

Variant Effect Predictor 

Gene expression in different tissues 

Find SNPs and other variants for my gene 

Retrieve gene sequence 

Compare genes across species 

Use my own data in Ensembl 

ENCODE data in Ensembl 

What's New in Ensembl Release 84 (March 2016)

- 20 haematopoietic primary cell epigenomes from the BLUEPRINT project
- Mouse: update to Ensembl-Havana GENCODE gene set
- Track hub registry interface
- dbSNP 146 for Human, Cow and Dog
- Pairwise LD calculation on LD variant page

[Full details](#) | [All web updates by release](#) | [More news on our blog](#)

• 02 Jun 2016: [What's coming in Ensembl release 85](#)

• 25 Apr 2016: [DNA day and Malaria day: a story of scientific endeavour](#)

• 31 Mar 2016: [Ensembl 85 and Ensembl Genomes 32](#)

[Go to Ensembl blog](#)

Tweets by @ensembl

e! Ensembl @ensembl Studying a dwarfism mutation in the PNKP gene #UsingEnsembl gene annotation buff.ly/24kBCE

e! Ensembl @ensembl .dzerbino is at the @KeystoneSymp on human variation talking about Ensembl functional annotation now buff.ly/1O4ySMK

[Embed](#) [View on Twitter](#)

 Ensembl is a joint project between EMBL-EBI and the Wellcome Trust Sanger Institute to develop a software system which produces and maintains automatic annotation on selected eukaryotic genomes.
www.ensembl.org/info/website/tutorials/orm37.html includes a list of additional current and previous funding bodies. How to cite Ensembl in your own publications.



Primer design

Good primer design is essential for successful reactions

1. Primer Length:

- 18-22 bp.
- Grande o suficiente para ter especificidade
- Curta o suficiente para se ligar facilmente ao template

Primer design

2. GC Content

- O número de G's e C's no primer como % no número total de bases, deve ser entre 40-60%

3. GC Clamp

- A presença de G ou C nas últimas 5 bases do extremo 3'
- Ajuda a promover a ligação específica, devido à ligação ser mais forte entre G e C
- Mais de 3 G's or C's devem ser evitados nas últimas 5 bases do extremo 3'

Primer design

4. Primer Melting Temperature (T_m)

- Por definição: temperatura à qual metade da cadeia se dissocia e fica “single stranded” e indica a estabilidade do duplex
- Primers com T_m entre 52-58 °C dão bons resultados
- Primers com $T_m > 65^{\circ}\text{C}$ têm tendência para “secondary annealing”
- O conteúdo GC indica a T_m

5. Primer Pair T_m Mismatch Calculation

- O par de primers deve ter uma T_m aproximada, para maximizar o produto de PCR.
- A diferença não deve exceder 5°C, ou pode levar à não amplificação

Primer design

6. Primer Annealing Temperature

- A temperature de melting é determinante para a temperatura de annealing
- Ta elevada produz pouca hidrilação entre Primer-Template, o que conduz a baixo produto de PCR
- Ta baixas levam a produtos inespecíficos, desencadeados por ligações erradas entre bases
- 3 a 5°C abaixo da Tm. Se for muito inespecífico, vamos subindo.

Primer design

7. Primer Secondary Structures

- Entre primers ou dentro do mesmo primer
- Levam a pouco ou nenhum produto de PCR
- Afectam annealing e consequentemente a amplificação
- Reduzem disponibilidade dos primers no mix

Primer design

6.1. Hairpins

- Dentro do primer
- Optimally a 3' end hairpin with a ΔG of -2 kcal/mol and an internal hairpin with a ΔG of -3 kcal/mol is tolerated generally.

ΔG

- Traduz a estabilidade do harpin
- A energia necessária para quebrar a estrutura secundária
- Valores negativos elevados: harpins estáveis, logo indesejáveis

Primer design

6.2. Self Dimer

- Entre 2 primers iguais (mesmo sentido)
- Produz menos produto de PCR, uma vez que há muitos mais primers disponíveis na amostra que de gene alvo
- Optimally a 3' end self dimer with a ΔG of -5 kcal/mol and an internal self dimer with a ΔG of -6 kcal/mol is tolerated generally.

6.3. Cross Dimer

- Entre o Fw e Rv
- Optimally a 3' end cross dimer with a ΔG of -5 kcal/mol and an internal cross dimer with a ΔG of -6 kcal/mol is tolerated generally.

7. 3' End Stability

- É o máximo ΔG das últimas 5 bases do extremo 3' end
- Um extremo 3' instável (ΔG menos negativo) resulta em “false priming”

Primer design

8. Repeats

- É um di-nucleótido que ocorre muitas vezes consecutivamente, ATATATAT
- Evitar pois leva a ligações erradas
- O número máximo é de 4 di-nucleotidos

9. Runs

- Evitar muitas bases iguais seguidas, AGC**GGGGGATGGGG**
- Levam a ligações erradas
- O nº máximo aceitável é de 4bp

Primer design

10. Evitar Template Secondary Structure

- Caso contrário a cadeia template impede a ligação dos primers

11. Avoid Cross Homology

- Primers apenas devem ligar-se a zonas únicas, caso contrário tornam-se inespecíficos
- Testar numa base de dados

12. Amplicon Length

- qPCR- ronda 100bp
- PCR – até 1000bp
- Determinado pela posição Fw e RV

Primer design

Primer Design using Software

- 1 Sequência

www.ncbi.nlm.nih.gov/tools/primer-blast/ ☆ :

Primer-BLAST A tool for finding specific primers

NCBI/ Primer-BLAST: Finding primers specific to your PCR template (using Primer3 and BLAST).

PCR Template

Enter accession, gi, or FASTA sequence (A refseq record is preferred) [?](#) [Clear](#)

Range

Forward primer From To [?](#) [Clear](#)

Reverse primer

Or, upload FASTA file Choose File No file chosen

Primer Parameters

Use my own forward primer (5'→3' on plus strand) [?](#) [Clear](#)

Use my own reverse primer (5'→3' on minus strand) [?](#) [Clear](#)

PCR product size
Min: 70 Max: 1000

of primers to return: 10

Primer melting temperatures (T_m)
Min: 57.0 Opt: 60.0 Max: 63.0 Max T_m difference: 3 [?](#)

Exon/intron selection

A refseq mRNA sequence as PCR template input is required for options in the section [?](#)

Exon junction span: No preference [?](#)

Exon junction match:
Exon at 5' side: 7 Exon at 3' side: 4

Minimal number of bases that must anneal to exons at the 5' or 3' side of the junction [?](#)

Primer pair must be separated by at least one intron on the corresponding genomic DNA [?](#)

Intron inclusion
Intron length range
Min: 1000 Max: 1000000 [?](#)

Primer design

Primer Design using Software

- Várias sequências (alinhamento)

Load Your Alignment File

No file chosen

Or paste your file below:

Select your file format

PrimaClade Options:

Max Num of Degeneracies Num of Align Gaps to Skip

Exclude Region Start Pos Exclude Region Length

Primer3 Options:

Primer Min Tm Opt Tm Max Tm

Primer GC% Min Opt Max

You can access the PrimaClade FAQ here: [PrimaClade FAQ](#)

[<- Back to the PrimaClade Home Page](#)

PrimaClade

Primer design

Primer Design using Software

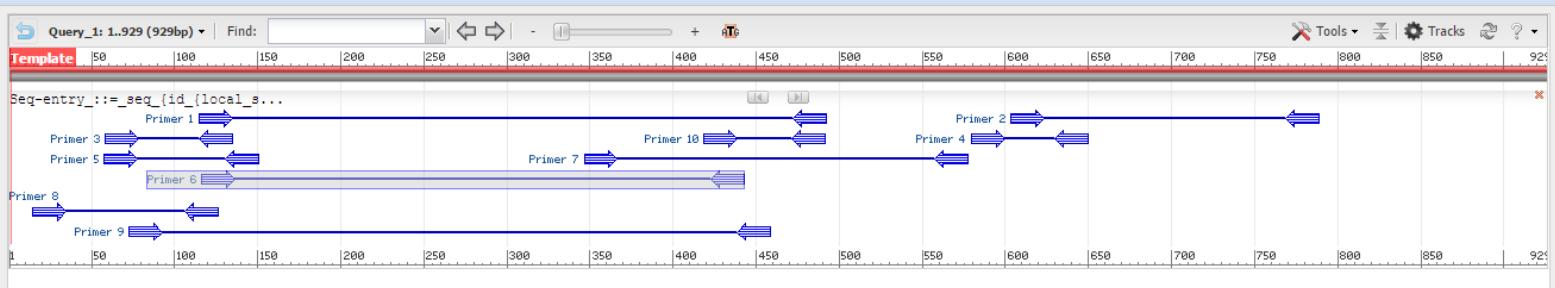
- PrimerBlast

Primer-BLAST *Primer-Blast results*

CBI/ Primer-BLAST : results: Job id=anGaZ-ltQEi7d395HldNBwV6fxYWZWIT [more...](#)

Input PCR template gi|365767550|gb|JQ247713.1| Heterocapsa nei isolate IFR10-193 large subunit ribosomal RNA gene, partial sequenceTACCGCTGAATATAAGCATATAAGTAAGCGGAGGATAAGAAACTAAATAGGATTCCCTCAGTAATGGCG
Range 1 - 929
Specificity of primers Primer pairs are specific to input template as no other targets were found in selected database: Refseq mRNA (Organism limited to Homo sapiens)
Other reports [► Search Summary](#)

Graphical view of primer pairs



Detailed primer reports

Primer pair 1								
	Sequence (5'->3')	Template strand	Length	Start	Stop	Tm	GC%	Self complementarity
Forward primer	GAGGGTGAGAGTCCCGTTG	Plus	20	115	134	60.04	60.00	3.00
Reverse primer	CTACCATGTCCTGGCGTTT	Minus	20	492	473	60.04	55.00	4.00
Product length	378							

Primer pair 2								
	Sequence (5'->3')	Template strand	Length	Start	Stop	Tm	GC%	Self complementarity
Forward primer	GAGGGTGAGAGTCCCGTTG	Plus	20	115	134	60.04	60.00	3.00
Reverse primer	CTACCATGTCCTGGCGTTT	Minus	20	492	473	60.04	55.00	4.00
Product length	378							

Primer design

Primer Design using Software

- Primaclade

primaclade.org/primaclade/EznOXqptnD.html

Generated primers for Untitled1.fas :

[Click here for your results in plain text](#)

gi 344310777 gb JN166011.11 gi 824555734 gb JAB987775.11 gi 824555731 gb JAB987772.11	-----1-----2-----3-----4-----5-----6-----7-----8-----9-----0-----1----- -----8-----9-----0-----1-----2-----3-----4-----5-----6-----7-----8----- -----TTGTCITTTAAATAAAAGACCTGTATGAATGCCAAAAACGAAAGTTC GTCCCCGCTGCCAGTGACAACATTAGTTCAACGGCGCGGTATTTGACCGTGCAAAGGTAGCGTAATCACTTGTCTTTAAATGAAGACCTGTATGAATGCCATAACGAGGGCTT GTCCCCGCTGCCAGTGACAACATTAGTTCAACGGCGCGGTATTTGACCGTGCAAAGGTAGCGTAATCACTTGTCTTTAAATGAAGACCTGTATGAATGCCATAACGAGGGCTT
Consensus	gtccccgctgccagtgacaacttagttcaacggccgcgttatttgacccgtgaaaggtagcgtaatcac TTGTCITTTAAATRAAGACCTGTATGAATGCCAWAACGARRGYT -----gcctggcccagtgacaact start pos=6 approx Tm=59.83 approx %gc=61.11 length=18 rev comp=agtgtcactggcaggc -----gtattttgacccgtcaaggtag start pos=40 approx Tm=59.95 approx %gc=43.48 length=19 -----ttgaccgtgcaaggtagc start pos=45 approx Tm=59.86 approx %gc=52.63 length=19 -----gaccgtgcaaggtagc start pos=47 approx Tm=60.41 approx %gc=61.11 length=19 -----ccgtgcaaggtagcgtaa start pos=49 approx Tm=59.88 approx %gc=52.63 length=19 -----ccgtgcaaggtagcgtaat start pos=49 approx Tm=60.15 approx %gc=50.00 1 -----gtgcaaggtagcgtaatcacTT start pos=51 approx Tm=59.74 approx %gc=43 -----caaaggtagcgtaatcacTTGTCTT start pos=54 approx Tm=60.12 approx %gc=43 -----aaaggtagcgtaatcacTTGTCTTTA start pos=55 approx Tm=60.02 approx %gc=43 -----aaaggtagcgtaatcacTTGTCTTTA start pos=55 approx Tm=60.03 approx %gc=43 -----aaggtagcgtaatcacTTGTCTTTA start pos=56 approx Tm=58.97 approx %gc=43 +++++TTAAATRAAGACCTGTATGAATGCC start pos=75 +++++TTAAATRAAGACCTGTATGAATGCC start pos=76 +++++TTAAATRAAGACCTGTATGAATGCC start pos=80 +++++TTAAATRAAGACCTGTATGAATGCCA start pos=81 +++++TRAAGACCTGTATGAATGCCAWA start pos=82 +++++AGACCTGTATGAATGCCAWAAGC start pos=83 +++++ACCTGTATGAATGCCAWAAGC start pos=84 +++++ACCTGTATGAATGCCAWAAGC start pos=85 ***** ***** ***** ***** ***** *****

Quick Output Key:

Lines like this ---- have 0 degenerate bases
 Lines like this +--+ have 1 to 2 degenerate bases
 Lines like this *-* have 3 or more degenerate bases

Primer design

Primer Design using Software

- Primaclade

gi 344310777 gb JN166011.1 gi 824555734 dbj AB987775.1 gi 824555731 dbj AB987772.1	-----TTGTCTTTAAATAAGACCTGTATGAATGGCAAA GTCCCCGCCTGCCAGTGACAACCTAGTTCAACGGCCGCGGTATTTGACCGTGCAAAGGTAGCGTAATCACTTGTCTTTAAATGAAGACCTGTATGAATGGCATAA GTCCCCGCCTGCCAGTGACAACCTAGTTCAACGGCCGCGGTATTTGACCGTGCAAAGGTAGCGTAATCACTTGTCTTTAAATGAAGACCTGTATGAATGGCATAA
Consensus	gtccccgacctggccaggtagacaacttagttcaacggccgcggtatttgaccgtgcaaaggtagcgtaatcac TTGTCTTTAAATRAAGACCTGTATGAATGGCAAWA -----gcctgcccaggtagacaact start pos=6 approx Tm=59.83 approx %gc=61.11 length=18 rev comp=agttgtcactggcaggc
	Fw Rev

Primer design

Primer Design using Software- analyse primers

<http://idtdna.com/calc/analyzer>

Sign In Get Help 0 Items € 0,00

Order Menu Products & Services ▾ Support & Education ▾ Tools ▾ Search Go

Achieve high efficiency qPCR under fast or standard cycling conditions
PrimeTime® Gene Expression Master Mix Try it for free »

OligoAnalyzer 3.1

Instructions | Definitions | Feedback

Sequence

5' - - 3'

0 Bases Parameter sets

SpecSheet (Default) ▾

Target type: DNA

Oligo Conc: 0.25 μM

Na⁺ Conc: 50 mM

Mg⁺⁺ Conc: 0 mM

dNTPs Conc: 0 mM

Analyze

Hairpin

Self-Dimer

Hetero-Dimer

NCBI Blast

Tm Mismatch

Results 5' Mods Internal Mods 3' Mods Mixed Bases

Standard Mixed Base Instructions

To use a Standard Mixed Base, simply type in the IUB symbol (from the table below) which represents the desired mix.

Custom Mixed Base Instructions

To use Custom Mixed Bases

Enter the desired percentage of each base (Integers Only, Totaling 100%).

Press 'Use Mix Base' button to add your custom mixed base.

Primer design

Primer Design using Software- analyse primers

custom oligos • next generation sequencing • qPCR • synthetic biology • RNAi • CRISPR genome editing [my order >>](#)

OligoAnalyzer 3.1 [Instructions](#) | [Definitions](#) | [Feedback](#)

Sequence: 20 Bases: 5'- ACC CGC TGA ATT TAA GCA TA -3' [Clear Sequence](#) [Add To Order](#)

Parameter sets: SpecSheet (Default)

Analyze: Hairpin

Target type: DNA
Oligo Conc: 0.25 μM
Na⁺ Conc: 50 mM
Mg⁺⁺ Conc: 0 mM
dNTPs Conc: 0 mM

Results: 5' Mods Internal Mods 3' Mods Mixed Bases

General Information: Batch date: 5/30/2016 3:05 PM

Nucleotide type: DNA Sequence type: Linear
Temperature: 25 °C Max Foldings: 20
Na Concentration: 50 mM Start Position: 0
Mg Concentration: 0 mM Stop Position: 0
Suboptimality: 50 % [Update](#) [Add To Order](#)

Structures:

Structure Name	Image	ΔG (kcal.mole ⁻¹)	T _m (°C)	ΔH (kcal.mole ⁻¹)	ΔS (cal.K ⁻¹ mole ⁻¹)	Output
1		-1.09	36.5	-29.4	-94.94	Ct Det

Length: 20bp (18-22)
 GC%: 40 (40-60)
 Hairpin: -1 (~-3)

Primer design

Primer Design using Software- analyse primers

Sequence: 5' ACCCGCTGAATTAAAGCATA -3'

Clear Sequence **Add To Order**

Target type: DNA
 Oligo Conc: 0.25 μM
 Na⁺ Conc: 50 mM
 Mg⁺⁺ Conc: 0 mM
 dNTPs Conc: 0 mM

Self-Dimer **Hetero-Dimer** **NCBI Blast** **Tm Mismatch**

Results **5' Mods** Internal Mods **3' Mods** Mixed Bases

Homo-Dimer Analysis

The delta G is calculated by taking into account the longest stretch of complementary bases. These pairs of complementary bases are represented by a solid line. Dotted lines represent additional complementary bases for that dimer structure, but their presence does not impact calculated delta G values. Actual delta G values may vary based on presence of additional complementary bases. The Maximum Delta G value refers to the free energy of the oligo sequence binding to its perfect complement.

Dimer Sequence:
 5'-ACCCGCTGAATTAAAGCATA
 : ||| :
 3' ATACGAATTAAAGTCGCCCA

Delta G: -5.36 kcal/mole Base Pairs: 4

5' ACCCGCTGAATTAAAGCATA
 : ||| :
 3' ATACGAATTAAAGTCGCCCA

Delta G: -4.85 kcal/mole Base Pairs: 4

5' ACCCGCTGAATTAAAGCATA
 : ||| :
 3' ATACGAATTAAAGTCGCCCA

Delta G: -4.74 kcal/mole Base Pairs: 3

5' ACCCGCTGAATTAAAGCATA
 ||| :: :: ::
 3' ATACGAATTAAAGTCGCCCA

GC%: 40 (40-60)
 Harpin: -1 (~-3)
 Self-dimer-: -5 (-5)

Primer design

Primer Design using Software- analyse primers

Instructions | Definitions | Feedback

Sequence

5'- ACCCGCTGAATTAAAGCATA -3'

20 Bases

Parameter sets

SpecSheet (Default)

Analyze

Hairpin

Self-Dimer

Hetero-Dimer

NCBI Blast

Tm Mismatch

Target type: DNA

Oligo Conc: 0.25 μM

Na⁺ Conc: 50 mM

Mg⁺⁺ Conc: 0 mM

dNTPs Conc: 0 mM

Results

5' Mods Internal Mods 3' Mods Mixed Bases

Hetero-Dimer Analysis

The delta G is calculated by taking into account the longest stretch of complementary bases. These pairs of complementary bases are represented by a solid line. Dotted lines represent additional complementary bases for that dimer structure, but their presence does not impact calculated delta G values. Actual delta G values may vary based on presence of additional complementary bases. The Maximum Delta G value refers to the free energy of the oligo sequence binding to its perfect complement.

Primary Sequence: 5'- ACCCGCTGAATTAAAGCATA -3'

Secondary Sequence: 5'- ACGAACGATTGACACGTAC -3'

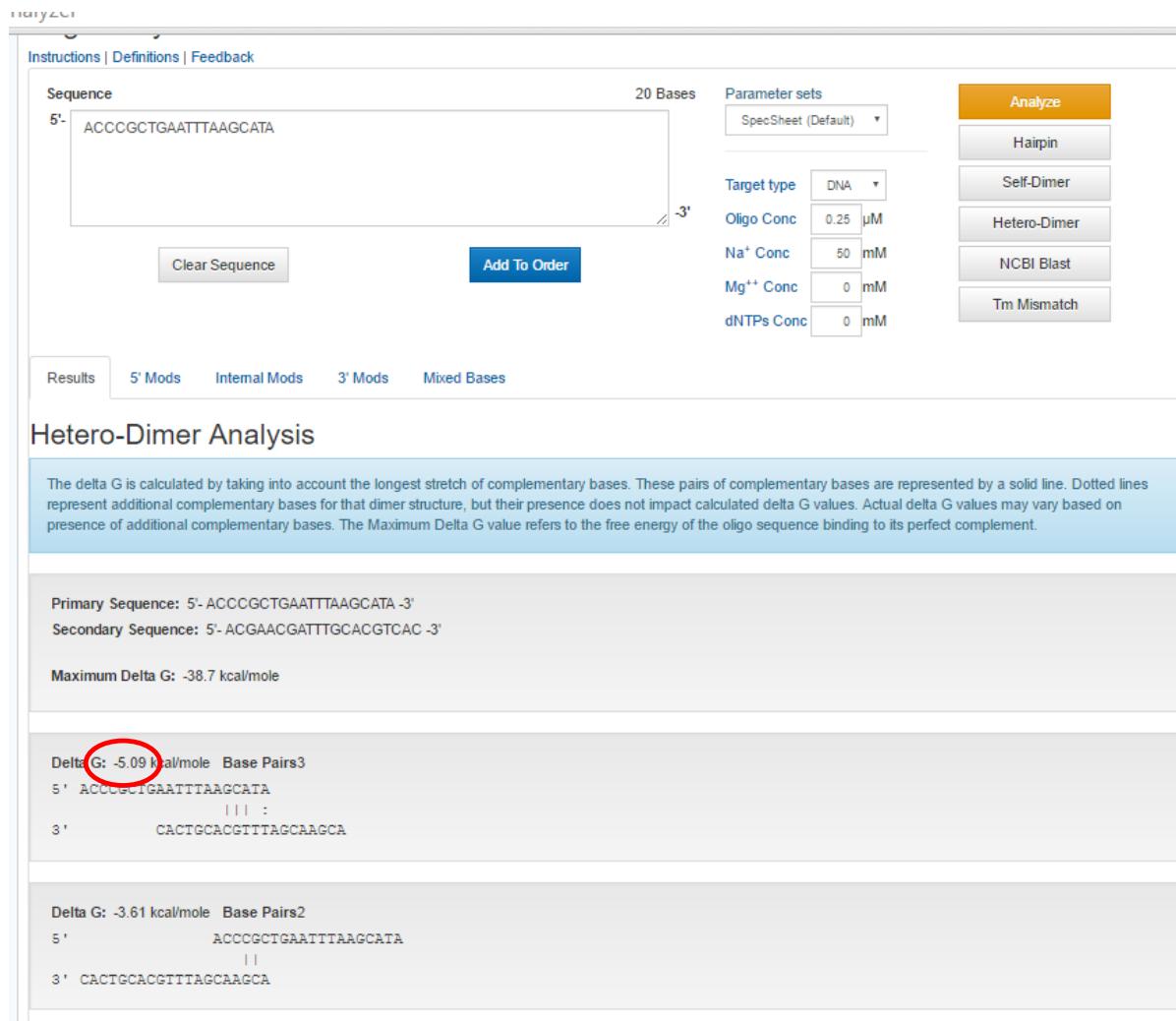
Maximum Delta G: -38.7 kcal/mole

Delta G: -5.09 kcal/mole Base Pairs3

5' ACCCGCTGAATTAAAGCATA
||| :
3' CACTGCACGTTAGCAAGCA

Delta G: -3.61 kcal/mole Base Pairs2

5' ACCCGCTGAATTAAAGCATA
|||
3' CACTGCACGTTAGCAAGCA



GC%: 40 (40-60)

Harpin: -1 (~-3)

Self-dimer: -5 (-5)

Hetero-dimer: -5 (-5)

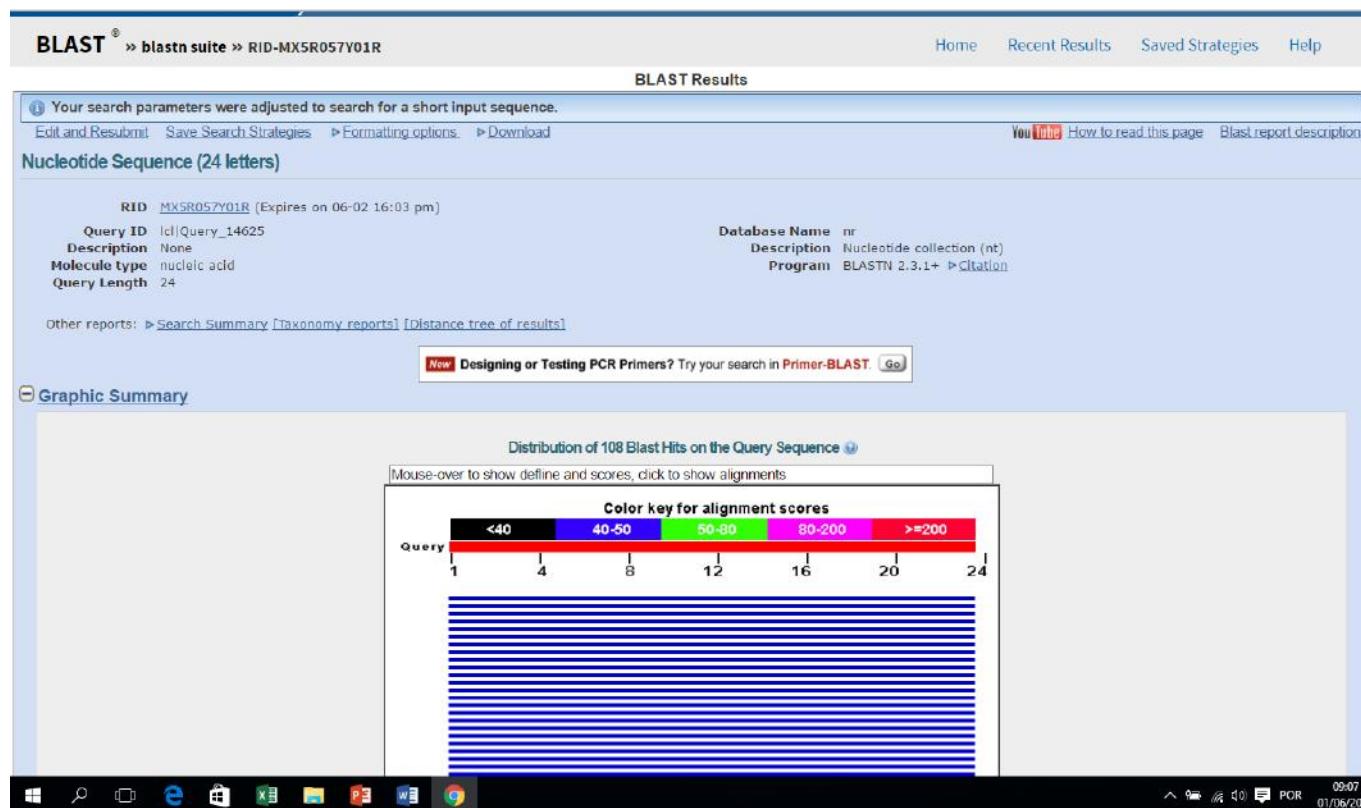
Tm (D1R): 52°C
(52-58 °C)

Tm (D3Ca): 56 °C
(diferença < 5°C)

Primer design

Primer Design using Software- analyse primers
(especificidade)

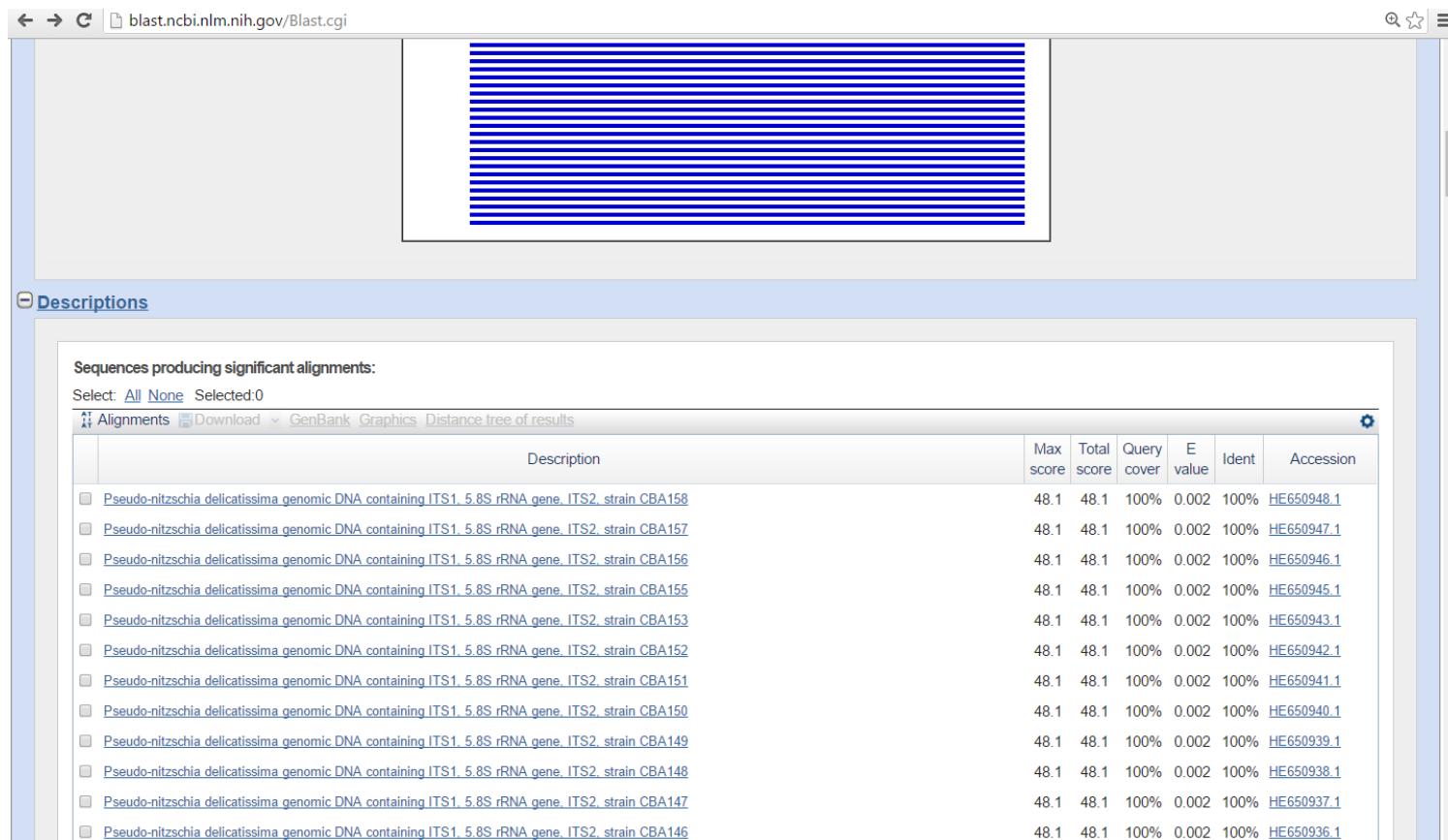
GTGCAATACTTGCTTGGGTTTCG



Primer design

Primer Design using Software- analyse primers
 GTGCAATACTTGCTGGGTTTCG

blast.ncbi.nlm.nih.gov/Blast.cgi



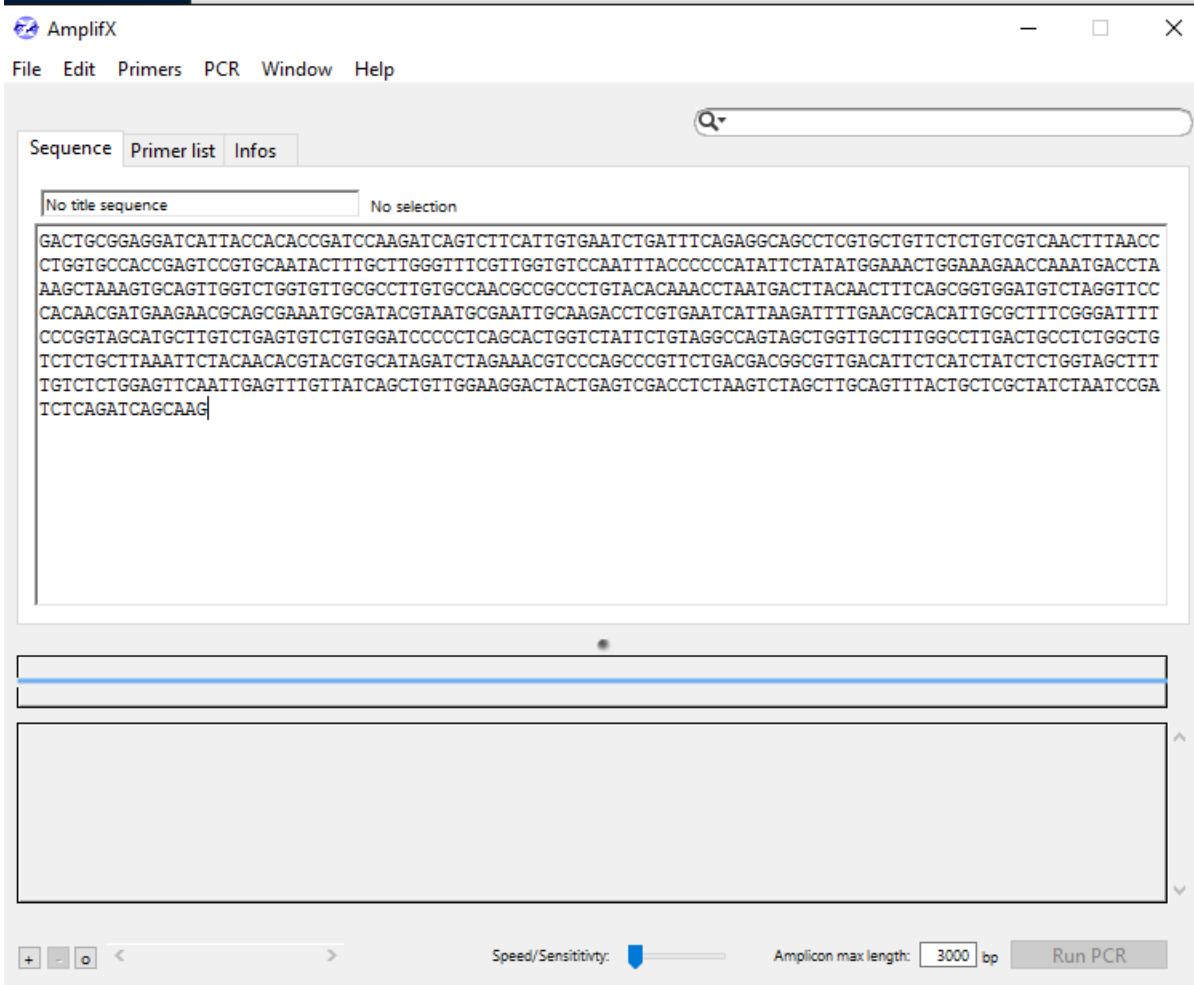
The screenshot shows a BLAST search results page from NCBI. At the top, there is a sequence alignment visualization consisting of several blue horizontal bars of varying lengths. Below this, the 'Descriptions' section is expanded, showing a table of sequences producing significant alignments. The table includes columns for Description, Max score, Total score, Query cover, E value, Ident, and Accession. All entries in the table are identical, listing 'Pseudo-nitzschia delicatissima genomic DNA containing ITS1_5.8S rRNA gene_ITS2_ strain CBA158' with a Max score of 48.1, Total score of 48.1, 100% Query cover, an E value of 0.002, 100% Ident, and Accession HE650948.1.

Description	Max score	Total score	Query cover	E value	Ident	Accession
Pseudo-nitzschia delicatissima genomic DNA containing ITS1_5.8S rRNA gene_ITS2_ strain CBA158	48.1	48.1	100%	0.002	100%	HE650948.1
Pseudo-nitzschia delicatissima genomic DNA containing ITS1_5.8S rRNA gene_ITS2_ strain CBA157	48.1	48.1	100%	0.002	100%	HE650947.1
Pseudo-nitzschia delicatissima genomic DNA containing ITS1_5.8S rRNA gene_ITS2_ strain CBA156	48.1	48.1	100%	0.002	100%	HE650946.1
Pseudo-nitzschia delicatissima genomic DNA containing ITS1_5.8S rRNA gene_ITS2_ strain CBA155	48.1	48.1	100%	0.002	100%	HE650945.1
Pseudo-nitzschia delicatissima genomic DNA containing ITS1_5.8S rRNA gene_ITS2_ strain CBA153	48.1	48.1	100%	0.002	100%	HE650943.1
Pseudo-nitzschia delicatissima genomic DNA containing ITS1_5.8S rRNA gene_ITS2_ strain CBA152	48.1	48.1	100%	0.002	100%	HE650942.1
Pseudo-nitzschia delicatissima genomic DNA containing ITS1_5.8S rRNA gene_ITS2_ strain CBA151	48.1	48.1	100%	0.002	100%	HE650941.1
Pseudo-nitzschia delicatissima genomic DNA containing ITS1_5.8S rRNA gene_ITS2_ strain CBA150	48.1	48.1	100%	0.002	100%	HE650940.1
Pseudo-nitzschia delicatissima genomic DNA containing ITS1_5.8S rRNA gene_ITS2_ strain CBA149	48.1	48.1	100%	0.002	100%	HE650939.1
Pseudo-nitzschia delicatissima genomic DNA containing ITS1_5.8S rRNA gene_ITS2_ strain CBA148	48.1	48.1	100%	0.002	100%	HE650938.1
Pseudo-nitzschia delicatissima genomic DNA containing ITS1_5.8S rRNA gene_ITS2_ strain CBA147	48.1	48.1	100%	0.002	100%	HE650937.1
Pseudo-nitzschia delicatissima genomic DNA containing ITS1_5.8S rRNA gene_ITS2_ strain CBA146	48.1	48.1	100%	0.002	100%	HE650936.1

Primer design

Primer Design using Software- analyse primers

Virtual PCR



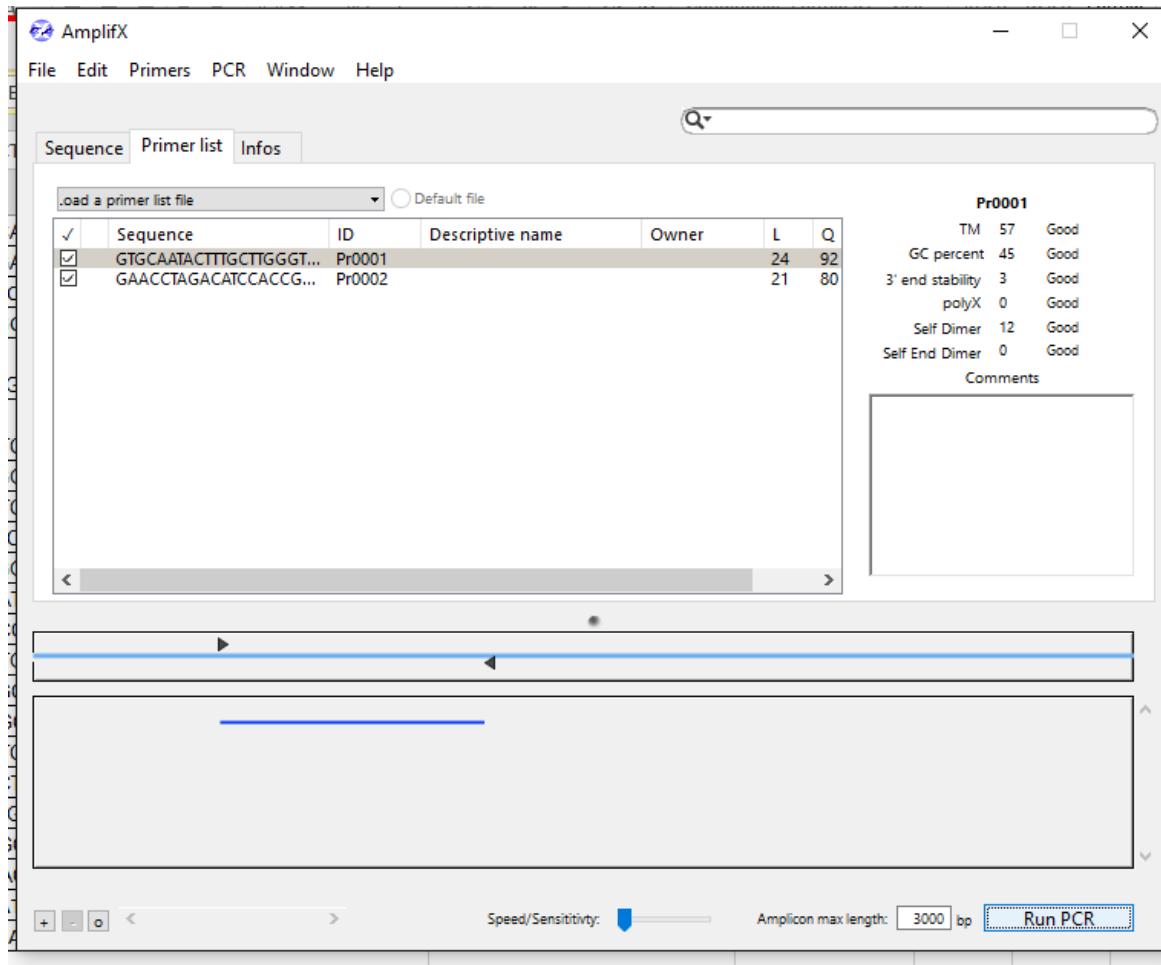
The screenshot shows the AmplifX software interface. The window title is "Virtual PCR". The menu bar includes File, Edit, Primers, PCR, Window, and Help. The tabs at the top are Sequence, Primer list, and Infos, with "Sequence" selected. The main area displays a sequence titled "No title sequence" with the ID "No selection". The sequence text is:

```
GAECTGCGGAGGGATCATTACCAACACCGATCCAAGAGTCAGTCTTCATTGTGAATCTGATTCAAGGGCAGCCCTCGTGCTGTTCTCTGCGTCAACTTTAACCTGTTGCCACCAGTCGCGACAATCTTGTGCTTGGGTTCTGGGTGTCCAATTACCCCCCATATTCTATATGGAAACTGGAAAGAACCAAATGACCTAAAGCTAAAGTGCAAGTGTGGAGTCAGTCTACAACTTAATGACTTACAACTTTCAGCGGTGGATGCTAGGTTCACAAACGATGAAGAACCGCAGCGAAATGCATACTGTAATGCGAATTGCAAGACCTCGTGAATCATTAAGATTTGAACGCACATTGCGCTTCGGGATTTCCCCTGAGCATGTTGCTGAGTGTCTGGATCCCCCTCAGGACTGGCTATTCTGTAGGCCAGTAGCTGGTTGACTGGCTCTGGCTGTCTCTGCTTAATCTACAACACGTACGTGCAAGATCTAGAAACGTCCTGACGACGGCGTTGACATTCTCATCTATCTGGTAGCTTTTGCTCTGGAGTICAATTGAGTTGTTATCAGCTGTTGGAGGACTACTGAGTCGACCTCTAAGTCTAGCTTACTGCTCGTATCTAATCCGATCTCAGATCAGCAAG
```

Below the sequence, there are two horizontal blue bars representing primers. At the bottom of the window, there are buttons for file operations (+, -, o, <, >), a slider for "Speed/Sensitivity", a text input for "Amplicon max length: 3000 bp", and a "Run PCR" button.

Primer design

Primer Design using Software- analyse primers
GTGCAATACTTGCTTGGGTTTCG



The screenshot shows the AmplifX software interface. The main window displays a sequence analysis for two primers:

Sequence	ID	Descriptive name	Owner	L	Q
GTGCAATACTTGCTTGGGTTTCG...	Pr0001			24	92
GAACCTAGACATCCACCG...	Pr0002			21	80

Primer Pr0001 details:

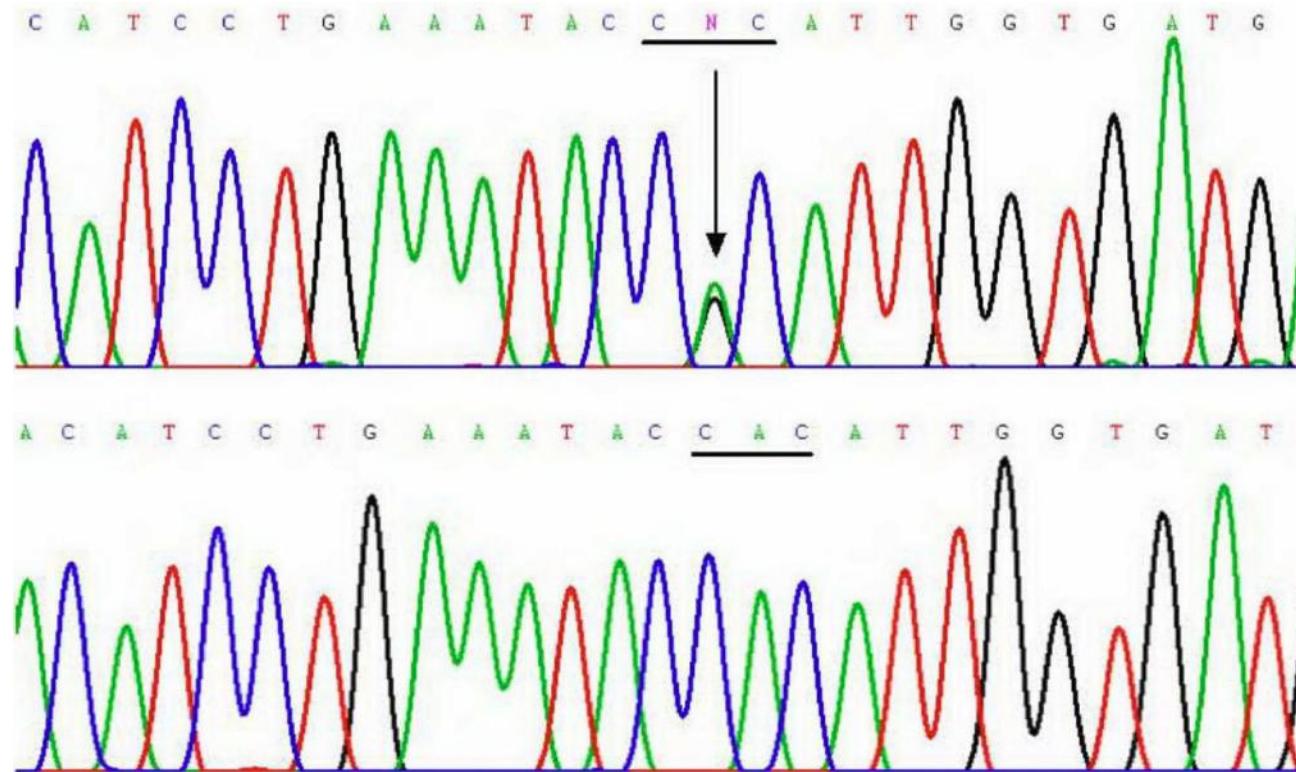
Parameter	Value	Status
TM	57	Good
GC percent	45	Good
3' end stability	3	Good
polyX	0	Good
Self Dimer	12	Good
Self End Dimer	0	Good

Comments: [Empty box]

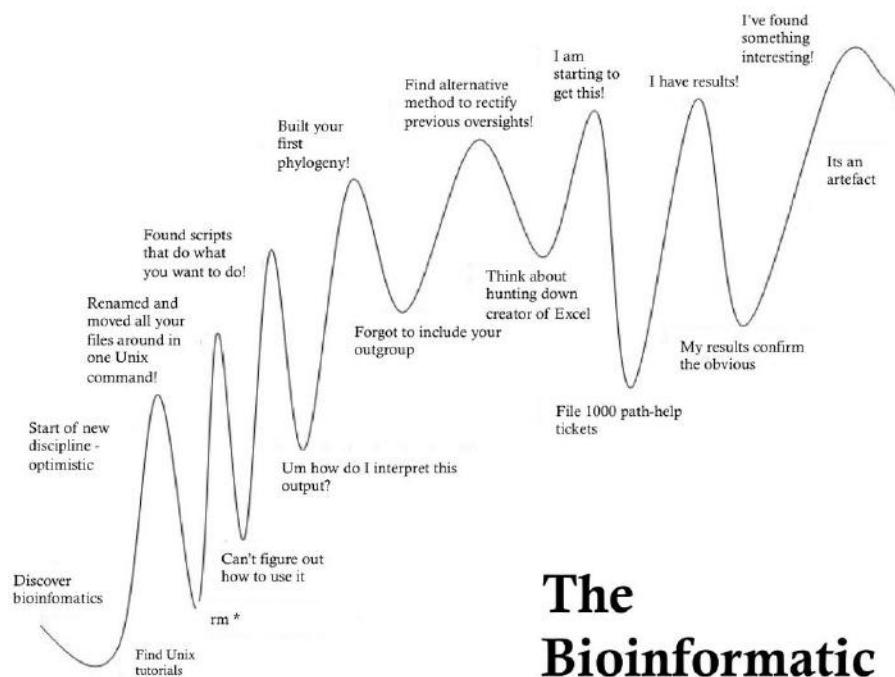
Below the table, there are two horizontal blue lines representing the primers, with arrows indicating their orientation. The top line corresponds to Pr0001 and the bottom line to Pr0002.

At the bottom of the interface, there are controls for Speed/Sensitivity (a slider), Amplicon max length (set to 3000 bp), and a Run PCR button.

Sequencing Results



Learning Curve



**The
Bioinformatic
learning curve**

Further reading

Swofford et al. 1996. Phylogenetic Inference. In Hillis, Moritz, & Mable [Eds.],

Molecular Systematics. Sinauer Associates, Sunderland, M.A.

Foster, 2007. Inferring phylogenetic relationships from sequence data. In Dear [Ed.],

Bioinformatics. Scion.

Lewis, 2001. Phylogenetic systematics turns of a new leaf. *Trends in Ecology and*

Evolution, 16: 30-37.

Felsenstein 2004. *Inferring Phylogenies*. Sinauer Associates, Sunderland, M.A.



Obrigada

Formadores
Bárbara Frazão
João Paulo Machado

Curso BIOMAR PT
2016





INTRODUÇÃO À BIOLOGIA MOLECULAR E BIOINFORMÁTICA: Bioinformatics

Lisboa, 12-14 October

João Machado

Bárbara Frazão

Bioinformatic

- What is it? Bioinformatics is the creation , development and operation of databases and other computational tools to collect, organize and interpret data
- Data Sources ? They are usually derived from biological data experiences that provide quantitative and qualitative data



Use of databases in bioinformatics as repositories and sources of information

Data Warehouses

- From 1982 databases began to be created for storing information and sequences of nucleotides

Examples

- European Molecular Biology Laboratory:
<http://www.embl.org/> (Europe)
- National Institutes of Health:
<http://www.ncbi.nlm.nih.gov> (North America)
- DNA Databank (DDBJ):
<http://www.ddbj.nig.ac.jp/> (Japan)

Data Warehouses

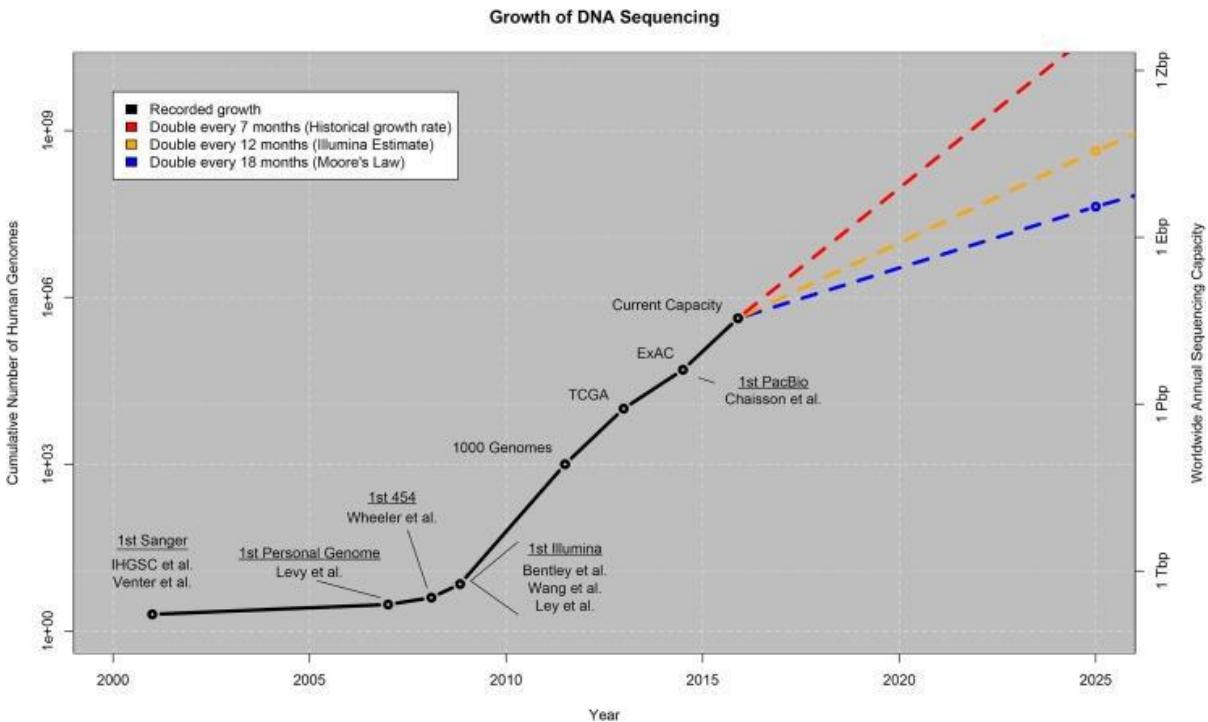
- From 1986 databases of amino acids (proteins)

Exemplos

- Swissprot/TrEMBL
- PIR
- In 2002 joined in UniProtKb
(<http://www.uniprot.org/>)
- **UniProtKB/Swiss-Prot** which is manually annotated and is reviewed and
- **UniProtKB/TrEMBL** which is automatically annotated and is not reviewed.

Data growth

- “100-gigabase” in August 2005. 200 billion bp in September 2007. The amount of data doubles every 18 months.



PLoS Biol. 2015 Jul; 13(7): e1002195.

Conventions

GenBank gb|accession.version

EMBL emb|accession.version

DDBJ dbj|accession.version

NCBI RefSeq ref|accession.version

PDB pdb|entry|chain

Patents pat|country|number

NBRF PIR pir| |entry

SWISS-PROT sp|accession|entry

Protein Research Foundation prf|name

Local Sequence identifier lcl|identifier

ref|NM_016519.5| Homo sapiens ameloblastin
(AMBN)

Conventions

RefSeq categories

Experimentally determined and curated		Genome annotation (computational predictions from DNA)	
NC	Complete genomic molecules		
NG	Incomplete genomic region		
NM	mRNA	XM	Model mRNA
NR	RNA (non-coding)		
NP	Protein	XP	Model protein

ref|NM_016519.5| Homo sapiens ameloblastin
(AMBN)

Conventions

IUPAC nucleotide ambiguity codes

Symbol	Meaning	Nucleic Acid
A	A	Adenine
C	C	Cytosine
G	G	Guanine
T	T	Thymine
U	U	Uracil
M	A or C	
R	A or G	
W	A or T	
S	C or G	
Y	C or T	
K	G or T	
V	A or C or G	
H	A or C or T	
D	A or G or T	
B	C or G or T	
X	G or A or T or C	
N	G or A or T or C	

Reference:

IUPAC-IUB SYMBOLS FOR NUCLEOTIDE NOMENCLATURE:
Cornish-Bowden (1985) *Nucl. Acids Res.* 13: 3021-3030.

Conventions

- Gene Names (<http://www.genenames.org/>)

HGNC
HUGO Gene Nomenclature Committee

Search everything ▾ Search symbols, keywords or IDs Use * to search with a root symbol (eg ZNF*)

Home Downloads Gene Families Tools Useful links About Newsletters Contact Us Help Request Symbol

HGNC is responsible for approving unique symbols and names for human loci, including protein coding genes, ncRNA genes and pseudogenes, to allow unambiguous scientific communication.

genenames.org is a curated online repository of HGNC-approved gene nomenclature, gene families and associated resources including links to genomic, proteomic and phenotypic information.

Search our catalogue of more than 39,000 symbol reports using our improved search engine (see [Search help](#)), search lists of symbols using our [Multi-symbol checker](#) and identify possible orthologs using our [HCOP](#) tool.

Download our ready-made data files from our [Statistics](#) and [Downloads](#) page, create your own datasets using either our [Custom Downloads](#) tool or [BioMart](#) service, or write a script/program utilising our [REST](#) service.

Submit your [gene symbol and name proposals](#) to us to be accredited with HGNC approved nomenclature for use in publications, databases and presentations.

FAQ

[What is the HGNC?](#)
[What is HGNC-approved nomenclature and why do we need it?](#)
[Where can I find information about existing human gene symbols?](#)
[What is a stem symbol?](#)
[Where can I find the Nomenclature Guidelines?](#)
[Do I have to use the approved symbols?](#)
[How should I cite HGNC nomenclature resources?](#)

Latest News

[Proposed change to the custom download tool \(give us your feedback\)](#)

We are proposing simplifying our "Custom Downloads" tool by bringing the data provided in line with that displayed in our symbol reports. Currently users can download two separate fields for some IDs: "HGNC curated" and "mapped data". This has caused some confusion as in our symbol reports HGNC curated data are displayed in preference, and mapped data are only shown if there is no HGNC curated ID i.e. only one ID is shown per symbol. Please use our [feedback](#) form to comment and let us know if this update may

File Formats

- Fasta files

```
>seq1
-----KSKERYKDENGNYFQLREDWW DANRE
>seq2
-----YEGLTTANGXKEYYQDKNGGNFFKLREDWW TANRE
>seq3
-----SQRHYKD-DGGNYFQLREDWW TANRH
>seq4
-----NVAALKTRYEK-DGQNFYQLREDWW TANYF
```

File Formats

- **Phylip interleaved**
- The first line of the input file contains the number of species and the number of characters separated by blanks. The information for each species follows, starting with a ten-character species name (which can include punctuation marks and blanks), and continuing with the characters for that species. Phylip format files can be interleaved, as in the example below, or sequential.

```
4 123
seq1 ----- ---KSKERYK DENGNYFQL REDWWDANRE
seq2 ----- YEGLT TANGXKEYYQ DKNGGNFFKL REDWWTANRE
seq3 ----- SQRHYK D-DGGNYFQL REDWWTANRH
seq4 ----- NVAALKTRYE K-DGQNFYQL REDWWTANRA

TVWKAITCNA --GGGKYFRN TCDG--GQNP TETQNNCRCIG-----
TVWKAITCGA P-GDASYFHA TCDSGDGRGG AQAPHKCRCG G-----
TVWEAITCSA DKGNA-YFRR TCNSADGKSQ SQARNQCRC- --KDENGKN-
TIWEAITCSA DKGNA-YFRA TCNSADGKSQ SQARNQCRC- --KDENGXN-
```

File Formats

- Phylip sequencial

4 123

```
seq1 ----- KSKERYK DENGNYFQL REDWWDANRE
TVWKAITCNA --GGKYFRN TCDG--GQNP TETQNNCRCI G-----
seq2 ----- YEGLT TANGXKEYYQ DKNGGNFFKL REDWWTANRE
TVWKAITCGA P-GDASYFHA TCDSGDGRGG AQAPHKCRCG G-----
seq3 ----- SQRHYK D-DGGNYFQL REDWWTANRH
TVWEAITCSA DKGNA-YFRR TCNSADGKSQ SQARNQCRC- --KDENGKN-
seq4 ----- NVAALKTRYE K-DGQNFYQL REDWWTANRA
TIWEAITCSA DKGNA-YFRA TCNSADGKSQ SQARNQCRC- --KDENGXN-
```

File Formats

- Nexus

```

#NEXUS
BEGIN DATA;
    DIMENSIONS NTAX=10 NCHAR=22;
    FORMAT MISSING=? DATATYPE=DNA GAP=- EQUATE="0=A 1=C";
    OPTIONS GAPMODE=MISSING;
MATRIX
[00000000000000000000]
[00000000000000000000]
[00000000011111111222]
[1234567890123456789012]

TaxonA          AAAAAAAAAAAAAAAA000000
TaxonB          AA----AAA--AAA1--100
TaxonC          AAA--AAAAAA--AAA010010
TaxonD          AAAGAA-AAAAGAA-A001001
TaxonE          ACGTACGTACGTACGT000000
TaxonF          AAAAAAAA000000
TaxonG          AA----AAA--AAA1--100
TaxonH          AAA--AAAAAA--AAA010010
TaxonI          AAAGAA-AAAAGAA-A001001
TaxonJ          ACGTACGTACGTACGT000000
;
END;

[ Indel Character      Sequence Region ]
[ -----           ----- ]
[                               ]
[ 17                 3-7           ]
[ 18                 4-5           ]
[ 19                 7-7           ]
[ 20                 11-12          ]
[ 21                 12-13          ]
[ 22                 15-15          ]

```

File Formats

GenBank

LOCUS	AF023787	618 bp	DNA	linear	PLN 02-MAY-1998	FEATURES	Location/Qualifiers
DEFINITION	Bryum stenotrichum small ribosomal protein 4 (rps4) gene, chloroplast gene encoding chloroplast protein, partial cds.					source	1..618 /organism="Bryum stenotrichum" /organelle="plastid:chloroplast" /mol_type="genomic DNA" /db_xref="taxon: 66994 "
ACCESSION	AF023787						
VERSION	AF023787.1	GI:3098167					
KEYWORDS	.					gene	<1..573 /gene="rps4"
SOURCE	chloroplast Bryum stenotrichum						
ORGANISM	Bryum stenotrichum					CDS	<1..573 /gene="rps4" /codon_start=1 /product="small ribosomal protein 4" /protein_id=" AAC15532.1 " /db_xref="GI:3098168" /translation="RRLGSLPGLTNKTPQLKNSINQSISNKKISQYRIRLEEKQKLR FHYGITERQLNYVRIARKAKGSTGEVLLQLEMRLDNVIFRLGMAPTIPGARQLVNH RHILVNDRIVNIPSYRCKPEDSITIKDRQKSQAIISKNLNLYQKYKTPNHLTYNFNLKK KGLVNQILDRESIGLKINELLVVEYYSRQA"
REFERENCE	1 (bases 1 to 618)						
AUTHORS	Cox,C.J. and Hedderson,T.A.J.						
TITLE	Phylogenetic relationships among the ciliate arthrodontous mosses: evidence from chloroplast and nuclear DNA sequences						
JOURNAL	Unpublished						
REFERENCE	2 (bases 1 to 618)						
AUTHORS	Cox,C.J. and Hedderson,T.A.J.						
TITLE	Direct Submission						
JOURNAL	Submitted (11-SEP-1997) Dept. of Botany, School of Plant Sciences, University of Reading, Whiteknights, Reading, Berkshire RG6 6AS, United Kingdom						
						ORIGIN	
							1 cgcgcgttag gatctttacc aggactaact aataaaacac cccagttaaa aactaattcg 61 atcaatcaat caatatctaa taaaaaaaatt tctcaatatc gcattcgtt ggaagaaaaa 121 caaaaattac gtttccatta tggataaca gagcgacaat tacttaatta tgtacgtatt 181 gctagaaaaag ctaaagggtc aacaggtgaa gtcttattac aattactga aatgcgttca 241 gataacgtta ttttcgatt aggtatggct cctacaattc ctggagcaag gcaactagta 301 aatcatagac atattttgta taatgatcgt atagtaataa taccatgtt tcgggtaaaa 361 cctgaggatt ctattactat aaaagatcga caaaaatctc aggctataat tagaaaaat 421 ttaaatttgt ataaaaata taaaacacca aatcattaa ttataattt tttaaaaaaa 481 aaaggatgg ttaatcaaact actagatcgt gaatccattt gttttaaaat aaatgattha 541 ttatgttag aatattatttc tcgccaagct taattaacaa ctaagagtgat ttgtatatt 601 atacataata aaaaatttg

//

File Conversions

- Software (e.g. Seaview, Mega, Mesquite, Bioedit, etc)
 - Requires instalation
- Scripts (generally in Python or Perl)
 - Several are freely distributed in github
(<https://github.com/>)
- Web-based tool (easy)

Hands-On 1

- <https://goo.gl/7DgwFF>
- 1) Descarregar as sequencias na pasta hands_on_1
 - 2) Usar o site: http://www.ebi.ac.uk/Tools/sfc/emboss_seqret/
 - 3) Analisar o formato original das sequencias
 - 4) Converter as sequencias para Fasta
 - 5) Adicionar num ficheiro separado usando editor de texto



Search Methods

Search NCBI

[NCBI](#) [Resources](#) [How To](#)

[Sign in to NCBI](#)

NCBI
National Center for
Biotechnology Information

All Databases ▾ [Search](#)

[NCBI Home](#)

[Resource List \(A-Z\)](#)

[All Resources](#)

[Chemicals & Bioassays](#)

[Data & Software](#)

[DNA & RNA](#)

[Domains & Structures](#)

[Genes & Expression](#)

[Genetics & Medicine](#)

[Genomes & Maps](#)

[Homology](#)

[Literature](#)

[Proteins](#)

[Sequence Analysis](#)

[Taxonomy](#)

[Training & Tutorials](#)

[Variation](#)

Welcome to NCBI

The National Center for Biotechnology Information advances science and health by providing access to biomedical and genomic information.

[About the NCBI](#) | [Mission](#) | [Organization](#) | [NCBI News](#) | [Blog](#)

Submit

Deposit data or manuscripts into NCBI databases



Download

Transfer NCBI data to your computer



Learn

Find help documents, attend a class or watch a tutorial



Develop

Use NCBI APIs and code libraries to build applications



Analyze

Identify an NCBI tool for your data analysis task



Research

Explore NCBI research and collaborative projects



Popular Resources

- [PubMed](#)
- [Bookshelf](#)
- [PubMed Central](#)
- [PubMed Health](#)
- [BLAST](#)
- [Nucleotide](#)
- [Genome](#)
- [SNP](#)
- [Gene](#)
- [Protein](#)
- [PubChem](#)

NCBI Announcements

- NCBI launches new Twitter account for NCBI Bookshelf 23 May 2016
- NCBI has a new Twitter feed - [@nchbooks](#) - to announce new books
- New NCBI Insights blog post: Fast Sequence Inspection with ORFfinder and SmartBLAST (PubMed Labs) 16 May 2016
- The latest book on NCBI Insights
- RefSeq release 76 is now available 18 May 2016
- RefSeq release 76 is accessible online, via FTP and through NCBI's programming utilities. This full release

[More...](#)

Search NCBI

NCBI Resources How To Sign in to NCBI

Search NCBI databases

Help

Results found in 32 databases for "pp2a"

Literature		Genes			
Books	102	books and reports	EST	1,681	expressed sequence tag sequences
MeSH	27	ontology used for PubMed indexing	Gene	18,715	collected information about gene loci
NLM Catalog	3	books, journals and more in the NLM Collections	GEO DataSets	32	functional genomics studies
PubMed	3,887	scientific & medical abstracts/citations	GEO Profiles	12,292	gene expression and molecular abundance profiles
PubMed Central	11,565	full-text journal articles	HomoloGene	87	homologous gene sets for selected organisms
Health		PopSet	35	sequence sets from phylogenetic and population studies	
ClinVar	6	human variations of clinical significance	UniGene	263	clusters of expressed transcripts
dbGaP	0	genotype/phenotype interaction studies	Proteins		
GTR	0	genetic testing registry	Conserved Domains	32	conserved protein domains
MedGen	7	medical genetics literature and links	Protein	110,089	protein sequences
OMIM	81	online mendelian inheritance in man	Protein Clusters	34	sequence similarity-based protein clusters
PubMed Health	0	clinical effectiveness, disease and drug reports	Structure	194	experimentally-determined biomolecular structures
Genomes		Chemicals			
Assembly	0	genome assembly information	BioSystems	6,046	molecular pathways with links to genes, proteins and chemicals
BioProject	25	biological projects providing data to NCBI	PubChem BioAssay	5,395	bioactivity screening studies
BioSample	0	descriptions of biological source materials	PubChem Compound	3	chemical information with structures, information and links
Clone	5	genomic and cDNA clones	PubChem Substance	178	deposited substance and chemical information
dbVar	995	genome structural variation studies			
Epigenomics	0	epigenomic studies and display tools			
Genome	2	genome sequencing projects by organism			
GSS	24	genome survey sequences			
Nucleotide	153,499	DNA and RNA sequences			
Probe	215	sequence-based probes and primers			
SNP	7,538	short genetic variations			
SRA	5	high-throughput DNA and RNA sequence read archive			
Taxonomy	0	taxonomic classification and nomenclature catalog			

Search NCBI

NCBI Resources How To Sign in to NCBI

Gene Gene pp2a Create alert Advanced Search

Gene sources Genomic Organelles Plasmids Categories Alternatively spliced Annotated genes Non-coding Protein-coding Pseudogene Sequence content CCDS Ensembl RefSeq RefSeqGene Status Current Chromosome locations more... Clear all Show additional filters

Did you mean pp2a as a gene symbol? Search Gene for pp2a as a symbol

Tabular 20 per page Sort by Relevance Send to: Hide sidebar >

Filters: Manage Filters

Results by taxon

Taxonomic Groups [List]

- eukaryotes (18226)
 - animals (7221)
 - chordates (5444)
 - arthropods (1385)
 - more... (441)
 - green plants (6249)
 - land plants (5981)
 - more... (329)
 - fungi (2516)
 - ascomycetes (1902)
 - more... (674)
 - apicomplexans (470)
 - oomycetes (285)
 - ciliates (253)
 - kinetoplastids (209)
 - Entamoeba (196)
 - cellular slime molds (80)
 - more... (616)
 - bacteria (425)
 - actinobacteria (133)
 - proteobacteria (124)
 - firmicutes (109)
 - more... (59)
 - archaea (55)
 - viruses (41)

Name/Gene ID	Description	Location	Aliases	MIM
<input checked="" type="checkbox"/> PP2A ID: 84333	serine/threonine protein phosphatase 2A [Arabidopsis thaliana (thale cress)]	Chromosome 1, NC_003070.9	AT1G69960, F20P5.30, F20P5_30, TYPE 2A SERINE/THREONINE PROTEIN PHOSPHATASE, serine/threonine protein phosphatase 2A	
<input checked="" type="checkbox"/> pp2A ID: 3878393	protein phosphatase 2A-like [Neurospora crassa OR74A]	Chromosome IV, NC_026504.1	(1637071..1639185, complement)	
<input checked="" type="checkbox"/> PP2A ID: 9680973	protein phosphatase 2A regulatory subunit [Micromonas pusilla COMP1545]		MICPUCDRAFT_30915	
<input checked="" type="checkbox"/> PPP2R4 ID: 5524	protein phosphatase 2 regulatory subunit 4 [Homo sapiens (human)]	Chromosome 9, NC_000009.12	PP2A, PR53, PTPA	600756
<input checked="" type="checkbox"/> Ppp2ca ID: 19052	protein phosphatase 2 (formerly 2A), catalytic subunit, alpha isform [Mus musculus (house mouse)]	Chromosome 11, NC_000077.6	PP2A, R75363	
<input checked="" type="checkbox"/> mts ID: 45059	microtubule star [Drosophila melanogaster]	Chromosome 2L, NT_022774.6	Dmel_CG7109_5569, CG7109, DmPp2A-2B, DmNCG7109, ER2-6, MTS/PP2A, Mts, PP2, PP2A, PP2A-2B, PP2A-C, PP2A-C, PP2A-MTS, PP2A-MT1, PP2A-MT1, PP2A-2B, PP2A-C, PP2A-2B, PP2A-C	

Find related data
Database: Select

Find items

Search details

pp2a[All Fields] AND alive[prop]

Search

See more...

Search NCBI

Gene sources
Genomic

Categories
Alternatively spliced
Annotated genes
 Protein-coding

Sequence content
 RefSeq

Status
 Current

Chromosome locations
more...

[Clear all](#)

[Show additional filters](#)

Tabular ▾ 20 per page ▾ Sort by Relevance ▾

Send to: ▾

[Hide sidebar >>](#)

Search results

Items: 1 to 20 of 67

Filters activated: Protein-coding, RefSeq. [Clear all](#) to show 57 items.

Showing Current items.

Name/Gene ID	Description	Location	Aliases
LOC100201603 ID: 100201603	serine/threonine-protein phosphatase 2A catalytic subunit beta isoform [<i>Hydra vulgaris</i>]		NEMVEDRAFT_v1g177129
NEMVEDRAFT_v1g177129 ID: 5522251	hypothetical protein [<i>Nematostella vectensis</i> (starlet sea anemone)]		NEMVEDRAFT_v1g175766
NEMVEDRAFT_v1g175766 ID: 5501708	hypothetical protein [<i>Nematostella vectensis</i> (starlet sea anemone)]		NEMVEDRAFT_v1g220857
NEMVEDRAFT_v1g220857 ID: 5500969	hypothetical protein [<i>Nematostella vectensis</i> (starlet sea anemone)]		NEMVEDRAFT_v1g195622
NEMVEDRAFT_v1g195622 ID: 5500404	hypothetical protein [<i>Nematostella vectensis</i> (starlet sea anemone)]		
LOC107353620 ID: 107353620	serine/threonine-protein phosphatase 2A activator-like [<i>Acropora digitifera</i>]		
LOC107352888 ID: 107352888	serine/threonine-protein phosphatase 2A catalytic subunit beta isoform [<i>Acropora digitifera</i>]		
LOC107349165 ID: 107349165	protein phosphatase 1H-like [<i>Acropora digitifera</i>]		

Find < Prev Page 1 of 3 Next > Last >>

Filters: [Manage Filters](#)

Results by taxon

Taxonomic Groups [List]

- cnidarians (57)
 - Anthozoa (39)
 - sea anemones (21)
 - stony corals (18)
 - hydromedusae (18)

Find related data

Database: [Select](#)

[Find aliases](#)

Search details

```
(pp2a[All Fields] AND "animals"[porgn] NOT "arthropods"
 [porgn] NOT "chordates"[porgn]) AND "cnidarians"[porgn] AND
 ("genotype protein coding"[Properties] AND "srcdb refseq"
 [Properties] AND alive[prop])
```

[Search](#)

[See more...](#)

Search NCBI

LOC107353620 serine/threonine-protein phosphatase 2A activator-like [Acropora digitifera]

Gene ID: 107353620, updated on 16-Apr-2016

Summary

Gene symbol: LOC107353620
 Gene description: serine/threonine-protein phosphatase 2A activator-like
 Gene type: protein coding
 RefSeq status: MODEL
 Organism: *Acropora digitifera*
 Lineage: Eukaryota; Metazoa; Cnidaria; Anthozoa; Hexacorallia; Scleractinia; Astrocoeniina; Acroporidae; Acropora

Genomic context

Location: chromosome: Un
 Exon count: 6

Annotation release	Status	Assembly	Chr	Location
100	current	Adig_1.1 (GCF_00022465.1)	Unplaced Scaffold	NW_015441583.1 (136387..141496)

See LOC107353620 in [Genome Data Viewer](#)

Genomic regions, transcripts, and products

Genomic Sequence: NW_015441583.1 Unplaced Scaffold Reference Adig_1.1 Primary Assembly

Go to reference sequence details
 Go to nucleotide: [Graphics](#) [FASTA](#) [GenBank](#)

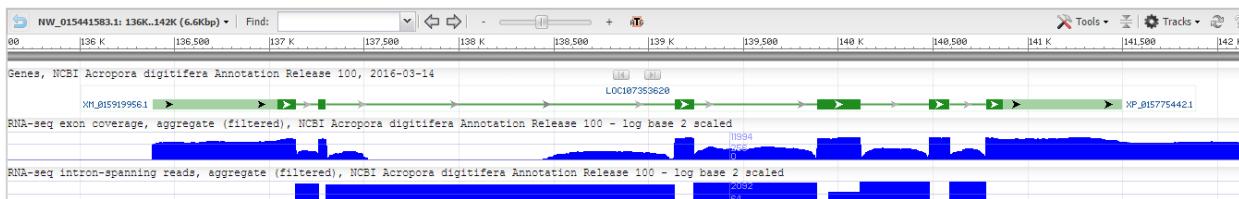


Table of contents

- Summary
- Genomic context
- Genomic regions, transcripts, and products
- Bibliography
- General protein information
- NCBI Reference Sequences (RefSeq)
- Related sequences

Genome Browsers

- Genome Data Viewer

Related information

- BioProjects
- Conserved Domains
- Gene neighbors
- Nucleotide
- Protein
- RefSeq Proteins
- RefSeq RNAs
- Taxonomy

General information

- About Gene
- FAQ
- FTP site
- Help
- My NCBI help
- NCBI Handbook
- Statistics

Related sites

- BLAST

Search NCBI

ⓘ NCBI is phasing out sequence GI numbers in September 2016. Please use accession.version! [Read more...](#)

GenBank ▾ Send: ▾ Change region shown ▾

PREDICTED: Acropora digitifera serine/threonine-protein phosphatase 2A activator-like (LOC107353620), mRNA

NCBI Reference Sequence: XM_015919956.1

[FASTA](#) [Graphics](#)

Go to: ▾

LOCUS	XM_015919956	1942 bp	mRNA	linear	INV 14-MAR-2016
DEFINITION	PREDICTED: Acropora digitifera serine/threonine-protein phosphatase 2A activator-like (LOC107353620), mRNA.				
ACCESSION	XM_015919956				
VERSION	XM_015919956.1 GI:1005477360				
DBLINK	BioProject: PRJNA314803				
KEYWORDS	RefSeq.				
SOURCE	Acropora digitifera				
ORGANISM	Acropora digitifera				
Eukaryota; Metazoa; Cnidaria; Anthozoa; Hexacorallia; Scleractinia; Astrocoeniina; Acroporidae; Acropora.					
COMMENT	<p>MODEL REFSEQ: This record is predicted by automated computational analysis. This record is derived from a genomic sequence (NW_015441583.1) annotated using gene prediction method: Gnomon, supported by mRNA and EST evidence.</p> <p>Also see:</p> <p>Documentation of NCBI's Annotation Process</p> <pre>##Genome-Annotation-Data-START## Annotation Provider :: NCBI Annotation Status :: Full annotation Annotation Version :: Acropora digitifera Annotation Release 100 Annotation Pipeline :: NCBI eukaryotic genome annotation pipeline Annotation Software Version :: 6.5 Annotation Method :: Best-placed RefSeq; Gnomon Features Annotated :: Gene; mRNA; CDS; ncRNA ##Genome-Annotation-Data-END##</pre>				
FEATURES	Location/Qualifiers				
source	1..1942 /organism="Acropora digitifera" /mol_type="mRNA"				

Customize view ▾

Analyze this sequence ▾

Run BLAST

Pick Primers

Highlight Sequence Features

Find in this Sequence

Reference sequence information ▾

RefSeq protein product

See the reference protein sequence for PREDICTED: serine/threonine-protein phosphatase 2A activator-like (XP_015775442.1).

More about the gene LOC107353620 ▾

LOC107353620 gene

Related information ▾

Annotated Genomic

BioProject

Gene

Protein

Taxonomy

Recent activity ▾

[Turn Off](#) [Clear](#)

Search NCBI

```

FEATURES          Location/Qualifiers
source           1..1942
/organism="Acropora digitifera"
/mol_type="mRNA"
/db_xref="taxon:70779"
/chromosome="Unknown"
/country="Japan:Okinawa, Kunigami, Oku"
1..1942
/gene="LOC107353620"
/note="Derived by automated computational analysis using
gene prediction method: Gnomon. Supporting evidence
includes similarity to: 6 mRNAs, 3 ESTs, 25 Proteins, and
100% coverage of the annotated genomic feature by RNAseq
alignments, including 25 samples with support for all
annotated introns"
/db_xref="GeneID:107353620"
659..1315
/gene="LOC107353620"
/codon_start=1
/product="serine/threonine-protein phosphatase 2A
activator-like"
/protein_id="XP_015775442.1"
/db_xref="GI:1005477361"
/db_xref="GeneID:107353620"
/translation="MPLQQSVHSLVQPLLPDKFLGAAIELTAYLKDAFGNKTRIDYGT
GHEASFAAFLCCLFKLRLVLDQSDCAAIVFKVFQRYLELMRRLQLTYRMEPAGSQGVWG
LDDFQFLPIWGSQALIGHTSLEPQHFTCEKNVEEHHNKMFLGCIRFINQMKRGPF
EHSNTLWGSISSVKTWEKVNSGLMKMYKAEVLSKFPVIQHFVFGTLMSIKEGETFKKPL
"
ORIGIN
1 tcttgagtagc tgttagtcctt cccagcattg actgaccctag gctgggtgaac aatgtttttt
61 tttttttttt tttttttcat ttgtttctaa gggttaggaga aactactatg aaaggctggg
121 aaaaatcaatt ttatccata cattaacttgc tcactatata agtatgccttgc tcactataaa
181 ctgaaaaatt aatgtgttatt cttttatccat cattttctttt tgttttttttt attttgcggaa
241 gtcctcaaaa aggggtgatat agcaaacact tttcttcattc acagcttgc cacatcccttgc
301 gtttattgttgc ctggccaaa agataggaaa atttttcttgc ctccaggagaa aatataata
361 ttgttgcattt ctcagaatgg cctgcttata atctggatggaa ttgtatcattt ctggtaattt
421 aggtattgttgc ttaatttttatt tggatggca attggatttt gcttgcattt gcttgcattt
481 aaaggccatgc attagaagag tccacatcat ttgggttgcag ttaaggtaattt tccatagttt

```

Send: ▾

Complete Record
 Coding Sequences
 Gene Features

Choose Destination

File Clipboard
 Collections Analysis Tool

Pick Primers

Send: ▾

Complete Record
 Coding Sequences
 Gene Features

Download features.

Format

FASTA Nucleotide ▾

Create File

Search NCBI

sequence.txt

```
1 >Id|XM_015919956.1_cds_XP_015775442.1_1 [gene=LOC107353620] [protein=serine/threonine-protein phosphatase 2A activator-like] [protein_id=XP_015775442.1] [location=659..1315]
2 ATGCCCTTCAACAGAGTGTCACTCGTTAGTCAACCTCTTACCCAGACAAGTTCTGGGCTGCCA
3 TTGAATTAACATGCATATTAAAAGATGCCCTTGGAAATAAAACAAGAACATAGACTATGGAACAGGTCTGCA
4 AGCTTCCTTGCTGCATTCTTTGTTTATTCAGCTCAGAGTATTGGACCAAAGTGAAGTGTGCTGCT
5 ATTGTGTTCAAGGTTTTCAAGGGTATTAGAACTGATGAGACGATTGAGCTCACTTACAGAACATGGAAC
6 CAGCTGGCAGTCAGGGTGTGGGGGCTGGATGATTTCAAGTCTCCCTTCATTGGGGAAAGTGTCA
7 GCTGATAGGCCATACAAGTCTAGAGGCCACAGCACTTCACTTGTGAAAAAAACGTAGAGGGAGCATATAAC
8 AAGTACATGTTCTGGGCTGCATCGTTTATAACCAATGAAAAGAGGACCCCTTGAGAACACATTCCA
9 ACACTTGTGGGAATAAGCTGTAAACATGGAAAAAGTAAACTCTGGTTGATGAAAATGTATAA
10 AGCTGAGGTTCTATCCAAGTCCCAGTCATTCAAGCATTTGTTGGTACATTAATGTCTATAAAGAA
11 GGAGAAACGTTAAAAAACCCCTGTAA
12
13
```

Search NCBI

Species Summary ▾ 20 per page ▾ Sort by Default order ▾

Molecule types Selected: 3

Items: 1 to 20 of 736

1. [Caenorhabditis elegans Probable serine/threonine-protein phosphatase \(paa-1\). partial mRNA](#)
1,773 bp linear mRNA
Accession: NM_065761.4 GI: 392894997
GenBank FASTA Graphics

2. [Trichinella spiralis serine/threonine-protein phosphatase PP2A regulator, mRNA, partial cds](#)
159 bp linear mRNA
Accession: XM_003366924.1 GI: 339263803
GenBank FASTA Graphics

3. [Trichinella spiralis protein phosphatase PP2A \(Tsp_14159\) mRNA, partial cds](#)
394 bp linear mRNA
Accession: XM_003368715.1 GI: 339259723
GenBank FASTA Graphics

4. [Loa loa protein phosphatase PP2A regulatory subunit \(LOAG_00611\) mRNA, complete cds](#)
1,974 bp linear mRNA
Accession: XM_003136151.1 GI: 312066287
GenBank FASTA Graphics

Send to: ▾ **Filters:** [Manage Filters](#)

Choose Destination

File Clipboard Collections

Download 3 items.

Format: **FASTA** ▾

Sort by: **Default order** ▾

[Create File](#)

Trichinidae [List] (6)

Trichuridae (26)

Spirurida (155)

Ascarididae (110)

Loa (58)

Brugidae (22)

Toxocaridae (21)

Diplogasteridae (15)

Tylenchidae (3)

Wuchereria (265)

Trichuridae (40)

Spirurida (96)

Loa (45)

Brugia (32)

Wuchereria (19)

Ascarididae (36)

Toxocaridae (19)

Ascandidae (17)

Diplogasteridae (12)

Tylenchida (1)

Find related data

Database: **Select** ▾

[Find items](#)

Search NCBI

Courier New | 11 | B
3 total sequences

Mode: Select / Slide Selection: 0 Position: Sequence Mask: None Numbering Mask: None Start ruler at: 1

MI

10 20 30 40 50 60 70 80 90 100 110 120 130 140 150 160

```
gi|339263803|ref|XM_003366924.1 GCGCGTTGAAAGGATGGGAGAAGGACGTGGAAAGGAAAGCCTCGAAATCATGGAGAAACTTAACTTGACAAACCTGGGATCATGATTGAAAGTTAAAGTTAGATCCCTGGTCACTGAAATAGCTGGTTAGCGGCAGTTGCAATAAGAAAATGA
gi|339259723|ref|XM_003368715.1 CCGATGTAATATCATGCGTTGAATTAAATCATGATGGTAAATTGCTGGCTACCGGTGATAAGGGCGACGCGATTGTAATATTCAACGTTGATCARAGCAATAAGCTGGTGAATGGTCATCGAAAGTCCGAAATACAAGAGTTAGCACATTCAAGAGTC
gi|312066287|ref|XM_003136151.1 ATGGAGTTGGAACATCTGATTAAGGGGGTTGGGATGGGTGGGTTAGTGGGGAGGGGAGTCACCGCATCAGGACGAGTTAGTACRGTGAAAGGAGGGTGAACCGTCAAGGATGAGGTAGTGCACTTGGAGAGAGGGCATTCCAGG
```

Courier New | 11 | B
3 total sequences

Mode: Select / Slide Selection: null Position: Sequence Mask: None Numbering Mask: None Start ruler at: 1

MI

10 20 30 40 50 60 70 80 90 100 110 120

```
gi|339263803|ref|XM_003366924.1 SPLERIATVVEETVVWEKAIVESLRTLIVDKPWHDLEVKLDPVVGQLAAVAIRK*
gi|339259723|ref|XM_003368715.1 PMYHALNLLIMMVNCWLEVIKADALYFENVIKAIISW*MVIEVLNTKCLAHSRVMNQNLI**NL*K*KKK*IVFDG*KGKT*LIFYFPLMIKQ*NYGK*LNVISGLMAVGICCTNTALPGY
gi|312066287|ref|XM_003136151.1 MEGFTIDLGGLGWVGLVGRESTESGRVSTVEEEGGGDPGQDEVVHWERGGIQVRTRLGYVVIMASLQAHEDTDNNLYPIAILIDELRNEDVQLRINSIRKLSTIALGVERTRGELIQ
```

Nucleotide Search

NCBI Resources How To Sign in to NCBI

Search NCBI databases

ambn Help

Results found in 25 databases for "ambn"

Literature			Genes		
Books	0	books and reports	EST	0	expressed sequence tag sequences
MeSH	5	ontology used for PubMed indexing	Gene	110	collected information about gene loci
NLM Catalog	0	books, journals and more in the NLM Collections	GEO DataSets	0	functional genomics studies
PubMed	190	scientific & medical abstracts/citations	GEO Profiles	2,920	gene expression and molecular abundance profiles
PubMed Central	233	full-text journal articles	HomoloGene	1	homologous gene sets for selected organisms
Health			PopSet	5	sequence sets from phylogenetic and population studies
ClinVar	10	human variations of clinical significance	UniGene	7	clusters of expressed transcripts
dbGaP	2	genotype/phenotype interaction studies	Proteins		
GTR	1	genetic testing registry	Conserved Domains	2	conserved protein domains
MedGen	1	medical genetics literature and links	Protein	3,160	protein sequences
OMIM	8	online mendelian inheritance in man	Protein Clusters	0	sequence similarity-based protein clusters
PubMed Health	0	clinical effectiveness, disease and drug reports	Structure	0	experimentally-determined biomolecular structures
Genomes			Chemicals		
Assembly	2	genome assembly information	BioSystems	39	molecular pathways with links to genes, proteins and chemicals
BioProject	0	biological projects providing data to NCBI	PubChem BioAssay	0	bioactivity screening studies
BioSample	0	descriptions of biological source materials	PubChem Compound	1	chemical information with structures, information and links
Clone	1,046	genomic and cDNA clones	PubChem Substance	91	deposited substance and chemical information
dbVar	110	genome structural variation studies			
Epigenomics	0	epigenomic studies and display tools			
Genome	3	genome sequencing projects by organism			
GSS	0	genome survey sequences			
Nucleotide	473	DNA and RNA sequences			
Probe	185	sequence-based probes and primers			
SNP	3,461	short genetic variations			
SRA	0	high-throughput DNA and RNA sequence read archive			
Taxonomy	0	taxonomic classification and nomenclature catalog			

You are here: NCBI > GQuery Write to the Help Desk

UniProt

UniProtKB ambn

Advanced Search

BLAST Align Retrieve/ID mapping Help Contact

About UniProtKB Basket

UniProtKB results

Filter by:

- Reviewed (5) Swiss-Prot
- Unreviewed (115) TrEMBL

Popular organisms

- Human (4)
- Mouse (3)
- Bovine (2)
- Rat (2)
- Pig (1)

Other organisms

Search terms

Filter "ambn" as: protein name (37)

View by

- Taxonomy
- Keywords
- Gene Ontology
- Enzyme class
- Pathway

UniRef

Your results in sequence clusters with

Entry	Entry name	Protein names	Gene names	Organism	Length
Q9NP70	AMBN_HUMAN	Ameloblastin	AMBN	Homo sapiens (Human)	447
Q55189	AMBN_MOUSE	Ameloblastin	Ambn	Mus musculus (Mouse)	407
Q28989	AMBN_PIG	Ameloblastin	AMBN	Sus scrofa (Pig)	421
Q62840	AMBN_RAT	Ameloblastin	Ambn	Rattus norvegicus (Rat)	422
Q5M8P3	Q5M8P3_MOUSE	Ambn protein	Ambn mCG_119077	Mus musculus (Mouse)	422
Q9XSX7	AMBN_BOVIN	Ameloblastin	AMBN	Bos taurus (Bovine)	392
Q546D7	Q546D7_HUMAN	Ameloblastin	AMBN	Homo sapiens (Human)	447
Q3B861	Q3B861_HUMAN	AMBN protein	AMBN	Homo sapiens (Human)	446
Q811C6	Q811C6_CAVPO	Ameloblastin	ambn AMBN	Cavia porcellus (Guinea pig)	423
Q811C5	Q811C5_CAVPO	Ameloblastin	ambn AMBN	Cavia porcellus (Guinea pig)	407
B1ACP5	B1ACP5_CAPMR	AMBN	AMBN	Caperea marginata (Pygmy right whale) (Balaena marginata)	155
B1ACP4	B1ACP4_BALAC	AMBN	AMBN	Balaenoptera acutorostrata (Common minke whale) (Balaena rostrata)	156
B1ACQ4	B1ACQ4_PECTA	AMBN	AMBN	Pecari tajacu (Collared peccary) (Tayassu tajacu)	155
B1ACP7	B1ACP7_DELLE	AMBN	AMBN	Delphinapterus leucas (Beluga whale)	155
B1ACQ7	B1ACQ7_TAPIN	AMBN	AMBN	Tapirus indicus (Asiatic tapir) (Malayan tapir)	155
B1ACQ5	B1ACQ5_CAMDR	AMBN	AMBN	Camelus dromedarius (Dromedary) (Arabian camel)	150

UniProt

PTM / Processingⁱ

Molecule processing

Feature key	Position(s)	Length	Description	Graphical view	Feature identifier	Actions
Signal peptide ⁱ	1 – 26	26	Sequence analysis			 Add  BLAST
Chain ⁱ	27 – 447	421	Ameloblastin		PRO_0000001192	 Add  BLAST

Amino acid modifications

Feature key	Position(s)	Length	Description	Graphical view	Feature identifier	Actions
Modified residue ⁱ	37 – 37	1	Hydroxyproline By similarity			
Modified residue ⁱ	43 – 43	1	Phosphoserine By similarity			
Glycosylation ⁱ	112 – 112	1	O-linked (GalNAc...) By similarity			

Keywords - PTMⁱ

Glycoprotein, Hydroxylation, Phosphoprotein

Proteomic databases

PaxDb ⁱ	Q9NP70.
PRIDE ⁱ	Q9NP70.

PTM databases

PhosphoSite ⁱ	Q9NP70.
--------------------------	---------

Expressionⁱ

Tissue specificityⁱ

Ameloblast-specific. Located at the Tomes processes of secretory ameloblasts and in the sheath space between rod-interrod enamel.

Gene expression databases

UniProt

Interactionⁱ

GO - Molecular functionⁱ

- growth factor activity 

Protein-protein interaction databases

BioGrid ⁱ	106756. 2 interactions.
STRING ⁱ	9606.ENSP00000313809.

Structureⁱ

3D structure databases

ProteinModelPortal ⁱ	Q9NP70.
ModBase ⁱ	Search...
MobiDB ⁱ	Search...

Family & Domainsⁱ

Domains and Repeats

Feature key	Position(s)	Length	Description	Graphical view	Feature identifier	Actions
Repeat ⁱ	189 – 201	13	1			 Add  BLAST
Repeat ⁱ	202 – 214	13	2			 Add  BLAST

Sequence similaritiesⁱ

Belongs to the ameloblastin family. 

Keywords - Domainⁱ

Repeat, Signal

1. Escolher um gene
 - Ameloblastin (AMBN)
 - Enamelin (ENAM)
 - Hemopexin (HPX)
2. Ir à página do NCBI e descarregar sequencias de 5 espécies

NCBI – BLAST (Basic Local Alignment Serach Tool)

BLAST Basic Local Alignment Search Tool

Home Recent Results Saved Strategies Help

► NCBI/ BLAST Home

BLAST finds regions of similarity between biological sequences. [more...](#)

New Aligning Multiple Protein Sequences? Try the COBALT Multiple Alignment Tool. [Go](#)

BLAST Assembled Genomes

Choose a species genome to search, or [list all genomic BLAST databases](#).

<input type="checkbox"/> Human	<input type="checkbox"/> Oryza sativa	<input type="checkbox"/> Gallus gallus
<input type="checkbox"/> Mouse	<input type="checkbox"/> Bos taurus	<input type="checkbox"/> Pan troglodytes
<input type="checkbox"/> Rat	<input type="checkbox"/> Danio rerio	<input type="checkbox"/> Microbes
<input type="checkbox"/> Arabidopsis thaliana	<input type="checkbox"/> Drosophila melanogaster	<input type="checkbox"/> Apis mellifera

Basic BLAST

Choose a BLAST program to run.

nucleotide blast	Search a nucleotide database using a nucleotide query <i>Algorithms:</i> blastn, megablast, discontiguous megablast
protein blast	Search protein database using a protein query <i>Algorithms:</i> blastp, psi-blast, phi-blast
blastx	Search protein database using a translated nucleotide query
tblastn	Search translated nucleotide database using a protein query
tblastx	Search translated nucleotide database using a translated nucleotide query

Specialized BLAST

Choose a type of specialized search (or database name in parentheses.)

Make specific primers with [Primer-BLAST](#)

News

[BLAST 2.2.23 release](#)
A new version of the stand-alone applications is available.
Mon, 22 Mar 2010 15:00:00 EST
[More BLAST news...](#)

Tip of the Day

[How to do Batch BLAST jobs.](#)
BLAST makes it easy to examine a large group of potential gene candidates.
[More tips...](#)

Blast Search

BLAST Basic Local Alignment Search Tool

Home Recent Results Saved Strategies Help

NCBI/ BLAST/ blastn suite

blastn blastp blastx tblastn tblastx

Enter Query Sequence

Enter accession number, gi, or FASTA sequence Query subrange

Or, upload file

Job Title Enter a descriptive title for your BLAST search

Align two or more sequences

Choose Search Set

Database Human genomic + transcript Mouse genomic + transcript Others (nr etc.):
Nucleotide collection (nr/nt)

Organism Optional Exclude

Exclude Optional Models (XM/XP) Uncultured/environmental sample sequences

Entrez Query Optional Enter an Entrez query to limit search

Program Selection

Optimize for Highly similar sequences (megablast)
 More dissimilar sequences (discontiguous megablast)
 Somewhat similar sequences (blastn)
 Choose a BLAST algorithm

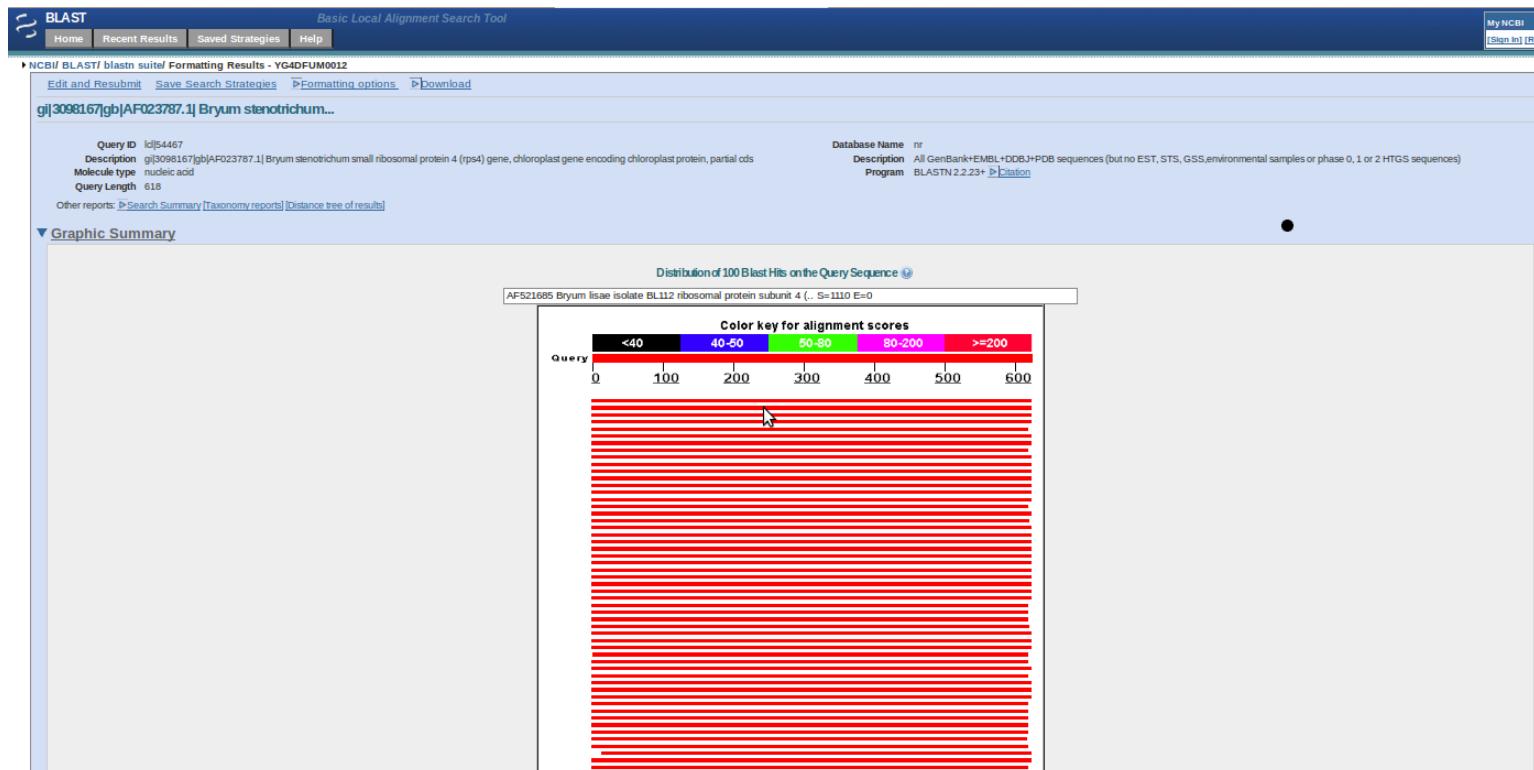
BLAST Search database Nucleotide collection (nr/nt) using Blastn (Optimize for somewhat similar sequences) Show results in a new window

Algorithm parameters Note: Parameter values that differ from the default are highlighted in yellow and marked with * sign

Blast methods

- blastn
- blastp
- blastx
- tblastn
- tblastx

NCBI – nucleotide BLAST result



NCBI – nucleotide BLAST result

BLAST Basic Local Alignment Search Tool

Home Recent Results Saved Strategies Help

NCBI/ BLAST/ blasts suite! Formatting Results - YG4DFUM0012

Edit and Resubmit Save Search Strategies ▶Formatting options ▶Download

gi|3098167|gb|AF023787.1| Bryum stenorhizum...

Query ID: gi|3098167|gb|AF023787.1| Bryum stenorhizum small ribosomal protein 4 (rp4) gene, chloroplast gene encoding chloroplast protein, partial cds
 Description: All GenBank+EMBL+DDBJ+PDB sequences (but no EST, STS, GSS environmental samples or phase 0, 1 or 2 HTGS sequences)
 Molecule type: nucleic acid
 Query Length: 618
 Program: BLASTN 2.2.23+ ▶ Citation

Other reports: ▶ Search Summary | Taxonomy reports | Distance tree of results

► Graphic Summary

▼ Descriptions

Legend for links to other resources: U UniGene E GEO G Gene S Structure M Map Viewer

Sequences producing significant alignments:

Accession	Description	Max score	Total score	Query coverage	E value	Max ident	Links
AF023787.1	Bryum stenorhizum small ribosomal protein 4 (rp4) gene, chloroplast gene encoding chloroplast protein, partial cds	1115	1115	100%	0.0	100%	
AF21087.1	Bryum lisae isolate BL112 ribosomal protein subunit 4 (rp4) gene, partial cds; chloroplast gene for chloroplast product	1110	1110	100%	0.0	99%	
AF521080.1	Bryum arachnogonium isolate BA107 ribosomal protein subunit 4 (rp4) gene, partial cds; chloroplast gene for chloroplast product	1110	1110	100%	0.0	99%	
AY082294.1	Bryum radicum small ribosomal subunit 4 (rp4) gene, partial cds; chloroplast gene for chloroplast product	1106	1106	100%	0.0	99%	
AY078333.1	Bryum paleaceum small ribosomal subunit protein 4 (rp4) gene, partial cds; chloroplast gene for chloroplast product	1103	1103	99%	0.0	99%	
AF521689.1	Bryum pseudotriquetrum isolate BP116 ribosomal protein subunit 4 (rp4) gene, partial cds; chloroplast gene for chloroplast product	1101	1101	100%	0.0	99%	
AF023785.1	Bryum donarium small ribosomal protein 4 (rp4) gene, chloroplast gene encoding chloroplast protein, partial cds	1087	1097	100%	0.0	99%	
AY078329.1	Bryum purpureascens small ribosomal subunit protein 4 (rp4) gene, partial cds; chloroplast gene for chloroplast product	1094	1094	99%	0.0	99%	
AF521678.1	Bryum algovicum isolate BA105 ribosomal protein subunit 4 (rp4) gene, partial cds; chloroplast gene for chloroplast product	1092	1092	100%	0.0	99%	
AY070771.1	Rosularium albolimbatum isolate MCP423 small ribosomal protein subunit 4 (rp4) gene, partial cds; and tRNA-Ser ger	1092	1092	100%	0.0	99%	
AF521692.1	Haplodontium reticulatum isolate HR119 ribosomal protein 4 (rp4) gene, partial cds; chloroplast gene for chloroplast product	1090	1090	100%	0.0	99%	
AF521673.1	Acidodontium heteroneuron isolate AH100 ribosomal protein subunit 4 (rp4) gene, partial cds; chloroplast gene for chloroplast product	1088	1088	100%	0.0	99%	
AY163087.1	Brachymenium preissiarum ribosomal protein subunit 4 gene, partial cds; chloroplast gene for chloroplast product	1088	1088	100%	0.0	99%	
AF521694.1	Bryum caucasicum isolate MH121 ribosomal protein subunit 4 (rp4) gene, partial cds; chloroplast gene for chloroplast product	1083	1083	100%	0.0	98%	
AF521682.1	Bryum capillare isolate BC109 ribosomal protein subunit 4 (rp4) gene, partial cds; chloroplast gene for chloroplast product	1083	1083	100%	0.0	98%	
AY078320.1	Brachymenium acuminatum small ribosomal subunit 4 (rp4) gene, partial cds; chloroplast gene for chloroplast product	1081	1081	99%	0.0	99%	
DQ794323.1	Bryum bicolor small ribosomal subunit 4 (rp4) gene, partial cds; chloroplast	1079	1079	100%	0.0	98%	
AF521687.1	Bryum pachytheca isolate EP114 ribosomal protein subunit 4 (rp4) gene, partial cds; chloroplast gene for chloroplast product	1079	1079	99%	0.0	98%	
AY070773.1	Plagiothecium zeri isolate MDP207 small ribosomal protein subunit 4 gene, partial cds; and tRNA-Ser gene, partial cds	1079	1079	100%	0.0	98%	
AY163091.1	Bryum orthothecium ribosomal protein subunit 4 gene, partial cds; chloroplast gene for chloroplast product	1079	1079	100%	0.0	98%	
AY163090.1	Bryum coronatum ribosomal protein subunit 4 gene, partial cds; chloroplast gene for chloroplast product	1079	1079	100%	0.0	98%	
AY163088.1	Bryum clavatum ribosomal protein subunit 4 gene, partial cds; chloroplast gene for chloroplast product	1079	1079	100%	0.0	98%	
AY163086.1	Brachymenium philonotula ribosomal protein subunit 4 gene, partial cds; chloroplast gene for chloroplast product	1079	1079	100%	0.0	98%	
AY082593.1	Bryum ruderae small ribosomal subunit 4 (rp4) gene, partial cds; chloroplast gene for chloroplast product	1079	1079	100%	0.0	98%	
AF251311.1	Mielichhoferia macrocarpa chloroplast partial rp4 gene for ribosomal protein, subunit 4	1079	1079	100%	0.0	98%	
AF023786.1	Anomobryum julaceum small ribosomal protein 4 (rp4) gene, chloroplast gene encoding chloroplast protein, partial cds	1077	1077	100%	0.0	98%	
AF521690.1	Bryum uliginosum isolate BU117 ribosomal protein subunit 4 (rp4) gene, partial cds; chloroplast gene for chloroplast product	1074	1074	100%	0.0	98%	
AF521676.1	Anomobryum conicum isolate AC103 ribosomal protein subunit 4 (rp4) gene, partial cds; chloroplast gene for chloroplast product	1074	1074	100%	0.0	98%	

E-values: 10^{-4} often considered good enough for an assumption of homology

Query	961	GGAGGTGCACAAGGCTCCCCTATGCCGGAGGCCAACCCAGACAATCTAGAAAACCCAGCT	1020
Sbjct	1062	GGAGGTGCACAAGGCTCCCCTATGCCGGAGGCCAACCCAGACAATCTAGAAAACCCAGCT	1121
Query	1021	TTCCTTACAGAGCTAGAACCTGCTCCCCACGCAGGGCTCCTTGCTCTCCCTAAGGATGAC	1080
Sbjct	1122	TTCCTTACAGAGCTAGAACCTGCTCCCCACGCAGGGCTCCTTGCTCTCCCTAAGGATGAC	1181
Query	1081	ATTCCCGGCCTGCCAAGGAGCCCTTCAGGGAAGATGAAGGGACTCCCCAGCGTCACCCCA	1140
Sbjct	1182	ATTCCCGGCCTGCCAAGGAGCCCTTCAGGGAAGATGAAGGGACTCCCCAGCGTCACCCCA	1241
Query	1141	GCAGCTGCTGACCCACTGATGACCCCTGAATTAGCTGATTTATAGGACCTACGATGCT	1200
Sbjct	1242	GCAGCTGCTGACCCACTGATGACCCCTGAATTAGCTGATTTATAGGACCTACGATGCT	1301
Query	1201	GACATGACCACATCCGTGGATTTCAGGAAAGCAACCATGGATACCAACGATGGCCCCA	1260
Sbjct	1302	GACATGACCACATCCGTGGATTTCAGGAAAGCAACCATGGATACCAACGATGGCCCCA	1361
Query	1261	AACTCTCTGCAAACATCCATGCCAGGAAACAAAGCCAGGAGGCCGAGATGATGCATGAC	1320
Sbjct	1362	AACTCTCTGCAAACATCCATGCCAGGAAACAAAGCCAGGAGGCCGAGATGATGCATGAC	1421
Query	1321	GCATGGCATTTCAGAGGCCCTG	1343
Sbjct	1422	GCATGGCATTTCAGAGGCCCTG	1444

Range 2: 633 to 671 [GenBank](#) [Graphics](#) ▼ Next Match ▲ Previous Match ⚡ First Match

Score	Expect	Identities	Gaps	Strand
53.6 bits(58)	0.011	35/39(90%)	0/39(0%)	Plus/Plus

Query	610	CTCCCAAGGATTGGATTTGCTGATCCACAAGGTTCAACA	648
Sbjct	633	CTCCCAAGGAGTAGATTTGCTGATCCACAAGGTTCCATCA	671

Range 3: 711 to 749 [GenBank](#) [Graphics](#) ▼ Next Match ▲ Previous Match ⚡ First Match

Score	Expect	Identities	Gaps	Strand
53.6 bits(58)	0.011	35/39(90%)	0/39(0%)	Plus/Plus

Query	532	CTCCCAAGGAGTAGATTTGCTGATCCACAAGGTTCCATCA	570
Sbjct	711	CTCCCAAGGATTGGATTTGCTGATCCACAAGGTTCAACA	749

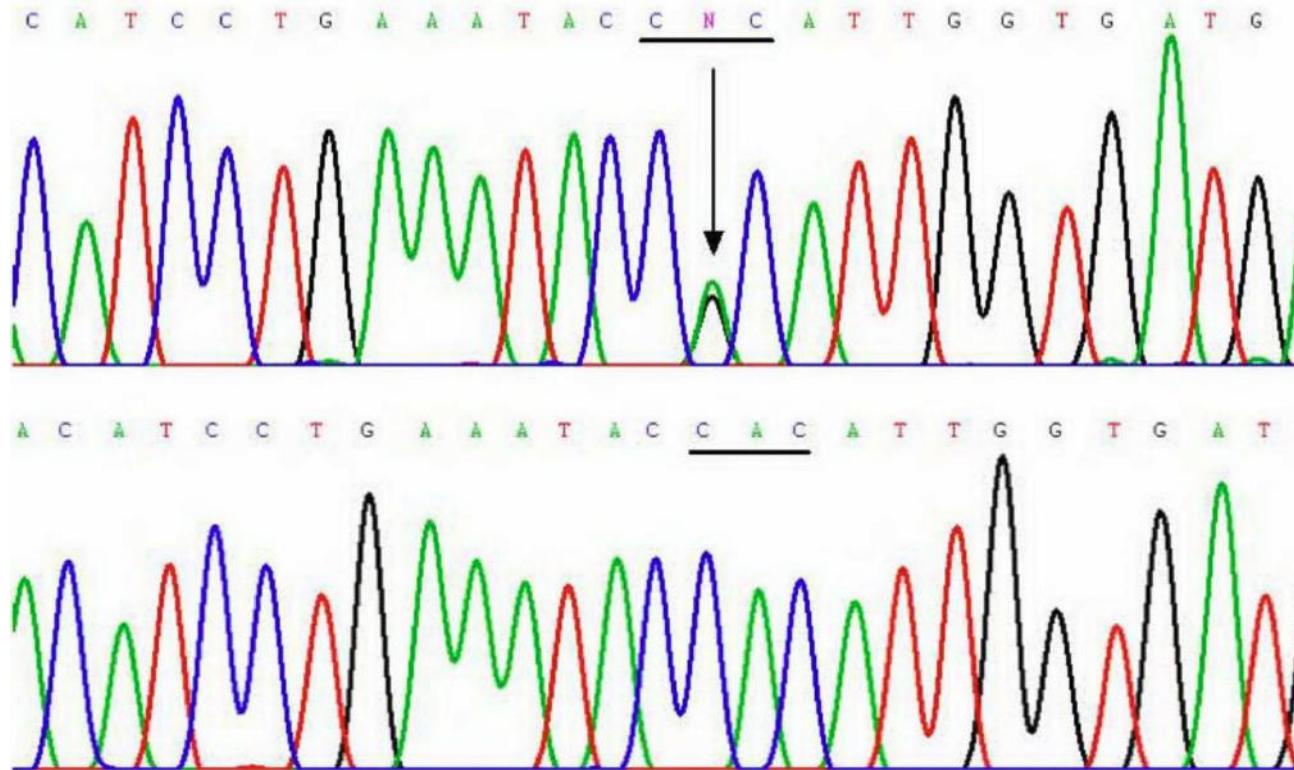
Warnings

- Predictions methods can fail and sometimes accuracy is not available
- Databases can contain incorrect data
- Avoid overvaloration of results

Hands-On 3

- 1) Descarregar sequencias da pasta hands_on_3
- 2) Fazer o “blast” para as sequencias
- 3) Identificar gene e espécies
- 4) Descarregar 3 top hits

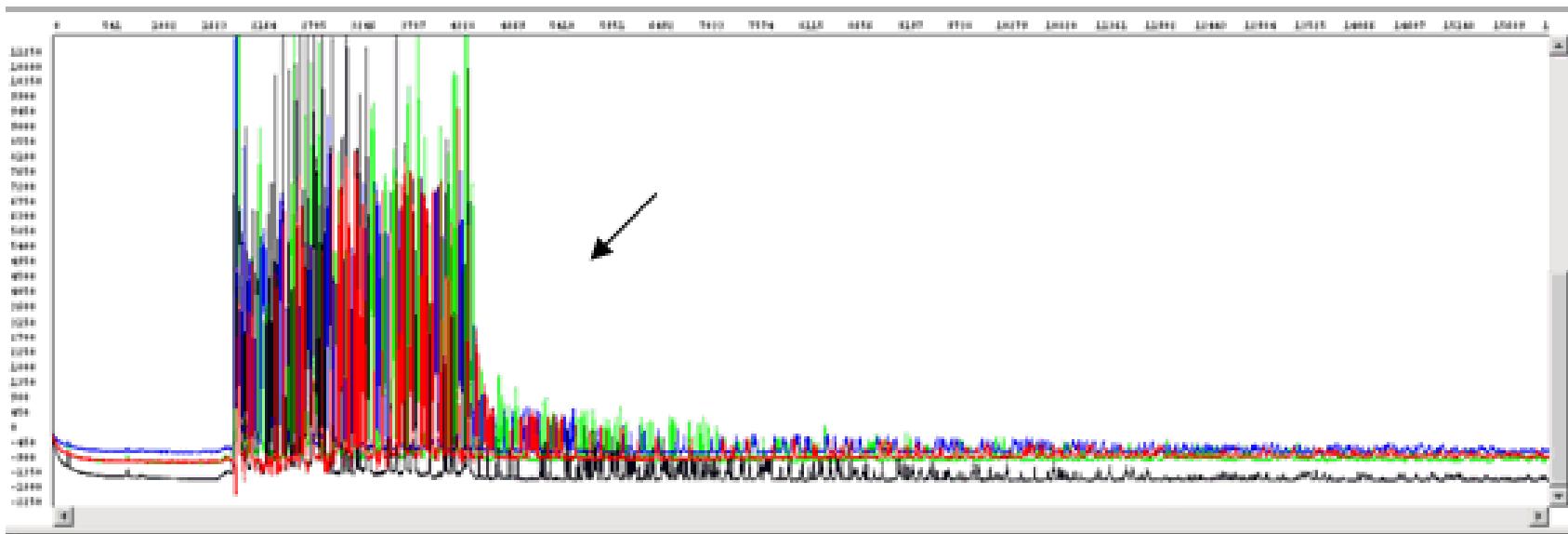
Sequencing Results



Sequencing Results



Sequencing Results



Sequencing Results

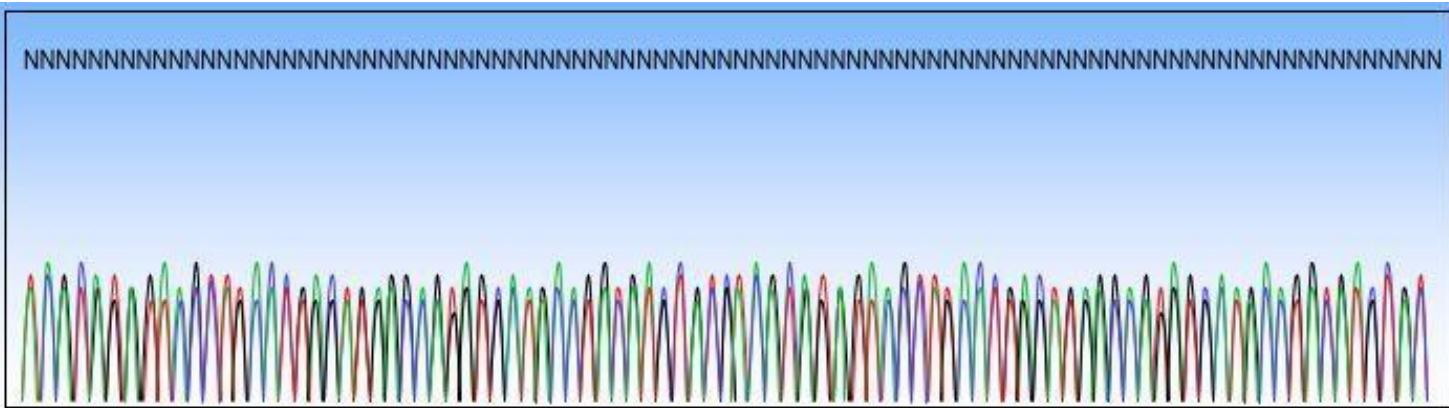
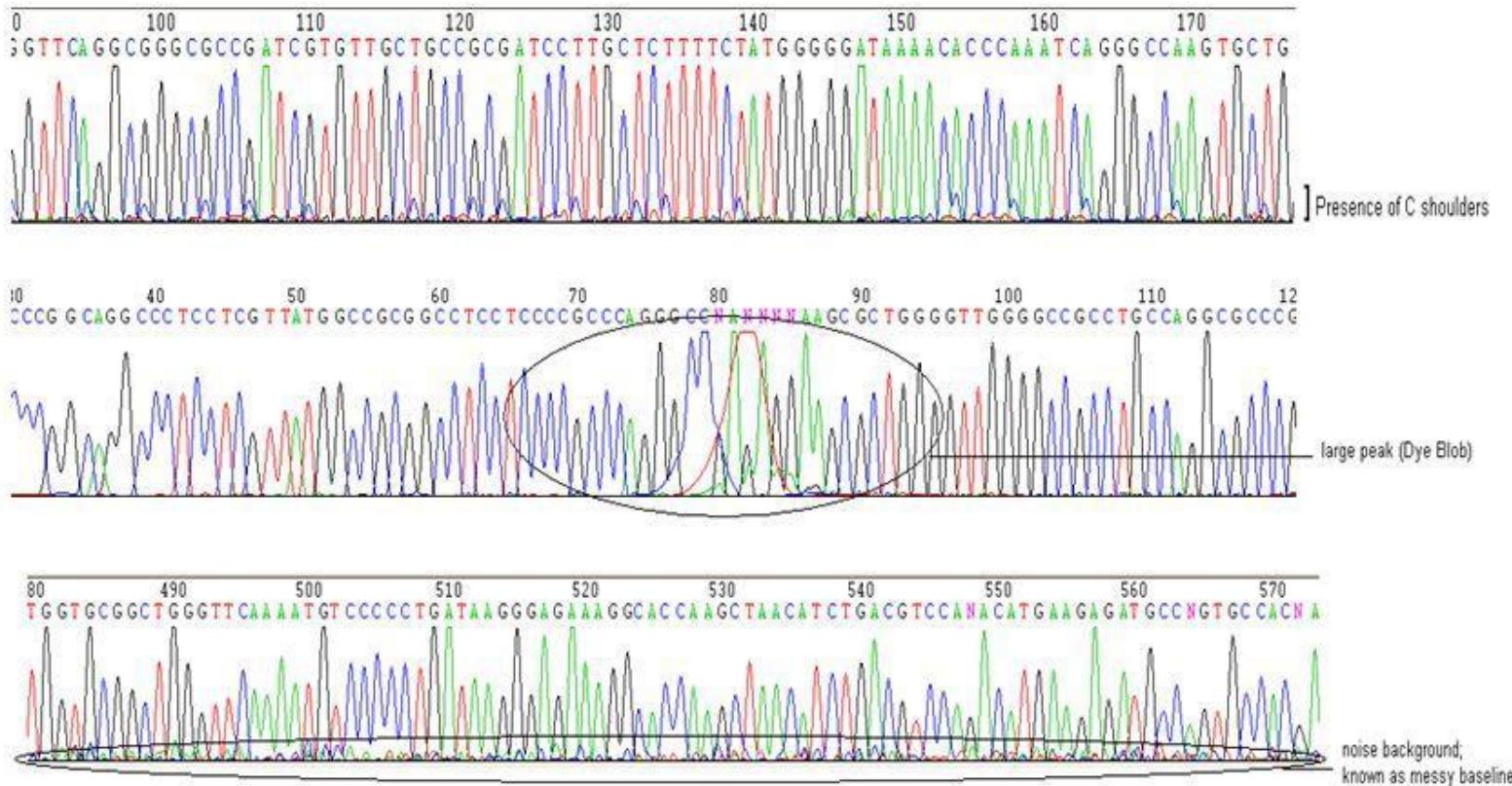


Figure 2: Double Priming Causes Double Sequence

A sample that has two annealing sights for a primer often show double sequence from the beginning of the result. In cases where both priming sights are closely located on the same strand of DNA, presence of the downstream primer could terminate extension of the upstream primer. The result is a single set of peaks located downstream, a positive indicator for double priming. The priming sights do not necessarily need to be identical as long as the 3 prime bases of the primer matches either annealing sight.

Sequencing Results





Methods for multiple sequence alignments

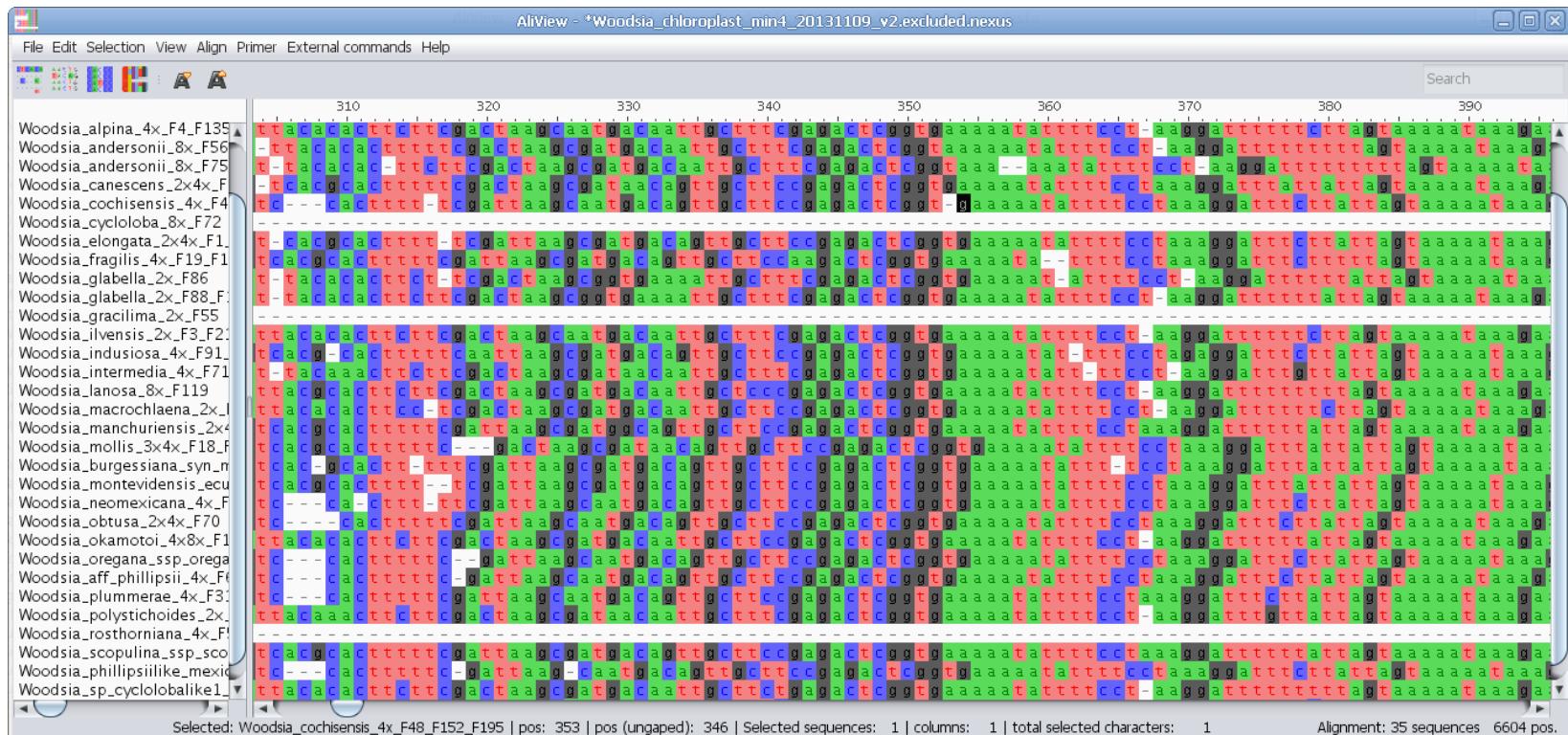
- Web-based (e.g. <http://translatorx.co.uk/>)
- Local (software)
- Scripts (e.g.
<http://raven.iab.alaska.edu/~ntakebay/teaching/programming/perl-scripts/perl-scripts.html>)

- Model
- Can be biased
- Quality can be checked (e.g.
<http://guidance.tau.ac.il/ver2/>)
- Parameters:
 - Gap penalties
 - Mismatch
 - Iterations
 - Guiding tree

MSA Algorithms

- ClustalW
- ClustalOmega
- Muscle
- T-Coffee
- MAFT
- Prank

MSA



MSA



FGFR1_10	MWSWKCLLFWAVLVITATLCTARPSPTIPEQ-	DALPSSSEDDDDDDDESSSEEKETDNTKPN--P	59
FGFR1_12	-	-	61
FGFR1_15	MWSWKCLLFWAVLVITATLCTARPSPTIPEQ-	DALPSSSEDDDDDDDESSSEEKETDNTKPNRMP	61
FGFR1_6	MWSWKCLLFWAVLVITATLCTARPSPTIPEQ-	DALPSSSEDDDDDDDESSSEEKETDNTKPNRMP	148
FGFR1_8	MWSWKCLLFWAVLVITATLCTARPSPTIPEQ-	DALPSSSEDDDDDDDESSSEEKETDNTKPNRMP	150
FGFR1_14	MWSWKCLLFWAVLVITATLCTARPSPTIPEQACPWGAPVVEVESFLVHPGDLLOLRCLRRLDDVQSINWL RDGVQLAESNRTRITGEEEVQDSVPADSGLYACVTSSPGSDTTYFSVNVDALPSSSEDDDDDDDESSSEEKETDNTKPNRMP	DALPSSSEDDDDDDDESSSEEKETDNTKPNRMP	150
FGFR1_4	MWSWKCLLFWAVLVITATLCTARPSPTIPEQACPWGAPVVEVESFLVHPGDLLOLRCLRRLDDVQSINWL RDGVQLAESNRTRITGEEEVQDSVPADSGLYACVTSSPGSDTTYFSVNVDALPSSSEDDDDDDDESSSEEKETDNTKPNRMP	DALPSSSEDDDDDDDESSSEEKETDNTKPNRMP	150
FGFR1	MWSWKCLLFWAVLVITATLCTARPSPTIPEQACPWGAPVVEVESFLVHPGDLLOLRCLRRLDDVQSINWL RDGVQLAESNRTRITGEEEVQDSVPADSGLYACVTSSPGSDTTYFSVNVDALPSSSEDDDDDDDESSSEEKETDNTKPNRMP	DALPSSSEDDDDDDDESSSEEKETDNTKPNRMP	150
FGFR1_11	-	-	61
FGFR1_13	-	-	61
FGFR1_7	MWSWKCLLFWAVLVITATLCTARPSPTIPEQ-	DALPSSSEDDDDDDDESSSEEKETDNTKPNRMP	61
FGFR1_9	MWSWKCLLFWAVLVITATLCTARPSPTIPEQ-	DALPSSSEDDDDDDDESSSEEKETDNTKPNRMP	61
FGFR1_17	MWSWKCLLFWAVLVITATLCTARPSPTIPEQ-	DALPSSSEDDDDDDDESSSEEKETDNTKPNRMP	61
FGFR1_18	MWSWKCLLFWAVLVITATLCTARPSPTIPEQ-	DALPSSSEDDDDDDDESSSEEKETDNTKPNRMP	59
FGFR1_2	MWSWKCLLFWAVLVITATLCTARPSPTIPEQACPWGAPVVEVESFLVHPGDLLOLRCLRRLDDVQSINWL RDGVQLAESNRTRITGEEEVQDSVPADSGLYACVTSSPGSDTTYFSVNVDALPSSSEDDDDDDDESSSEEKETDNTKPNRMP	DALPSSSEDDDDDDDESSSEEKETDNTKPNRMP	150
FGFR1_3	MWSWKCLLFWAVLVITATLCTARPSPTIPEQACPWGAPVVEVESFLVHPGDLLOLRCLRRLDDVQSINWL RDGVQLAESNRTRITGEEEVQDSVPADSGLYACVTSSPGSDTTYFSVNVDALPSSSEDDDDDDDESSSEEKETDNTKPNRMP	DALPSSSEDDDDDDDESSSEEKETDNTKPNRMP	150
FGFR1_16	MWSWKCLLFWAVLVITATLCTARPSPTIPEQAC-	DALPSSSEDDDDDDDESSSEEKETDNTKPNRMP	31
rule110.....20.....30.....40.....50.....60.....70.....80.....90.....100.....110.....120.....130.....140.....15032.....	32

Hands-On 4

- 1) Requisitos: Seaview e/ou Mega6
- 2) Descarregar sequencias da pasta Hands_on_4
- 3) Importar ficheiro para o Mega ou Seaview
- 4) Traduzir para aminoácidos
- 5) Alinhar usando Muscle e ou ClustalW
- 6) Filtrar resultados alinhamento usando o <http://translatorx.co.uk/>)

Q & A



INTRODUÇÃO À BIOLOGIA MOLECULAR E BIOINFORMÁTICA: Bioinformatics 2

Lisboa, 12-14 October

João Machado

Bárbara Frazão



Phylogenetic trees

Check List

- Recognize 5 file formats
- Perform a gene oriented search
- Perform a blast search
- Perform a MSA

The tree of Life

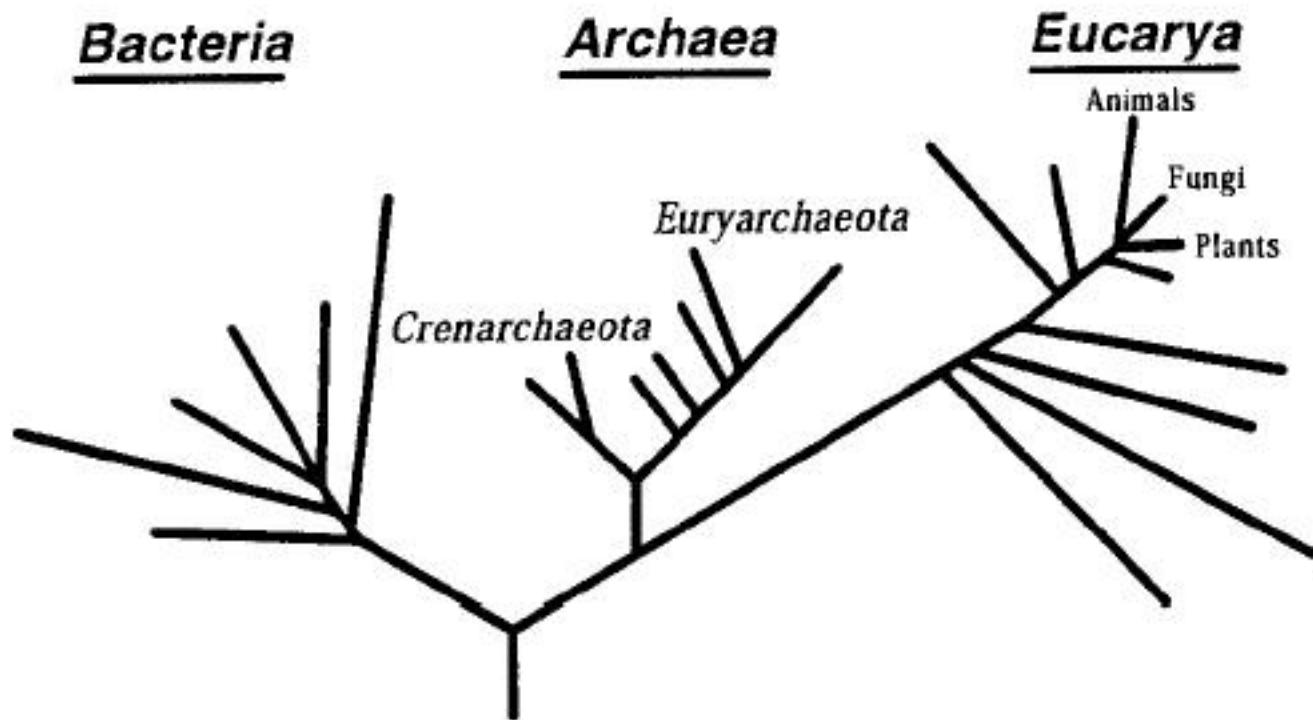
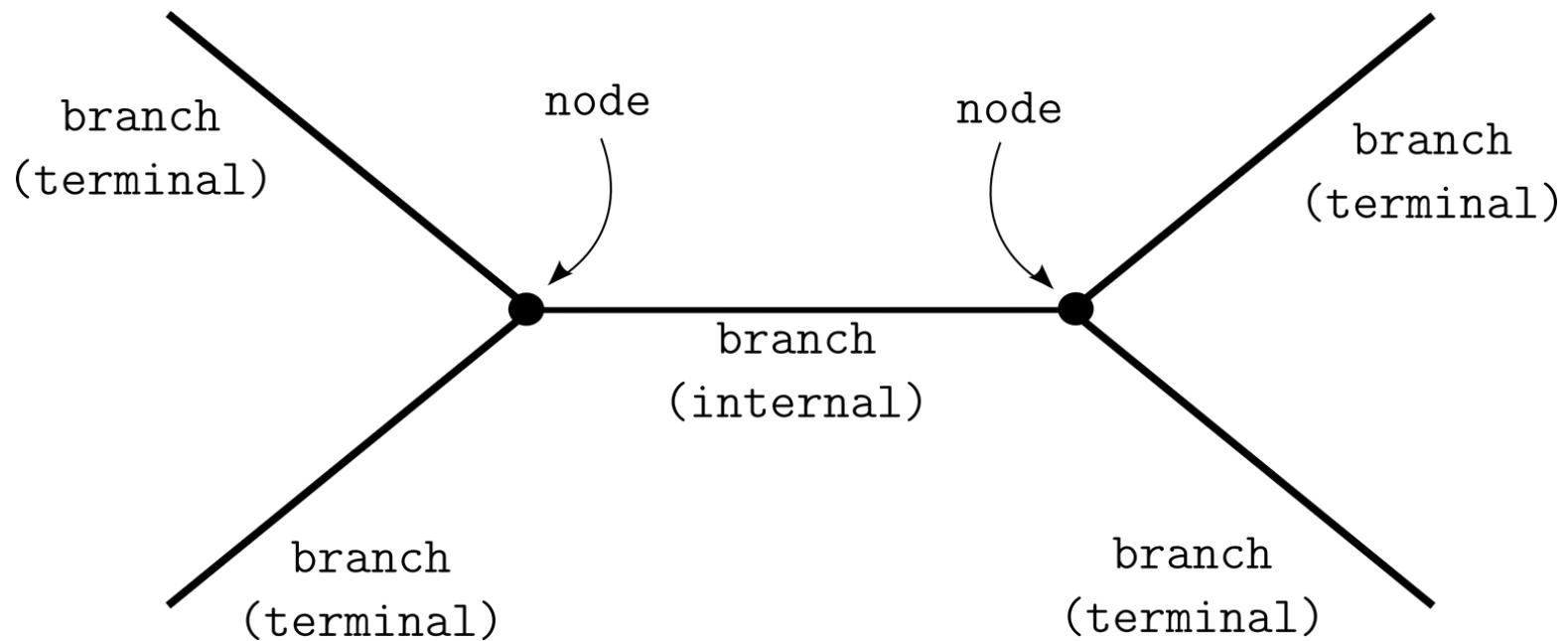
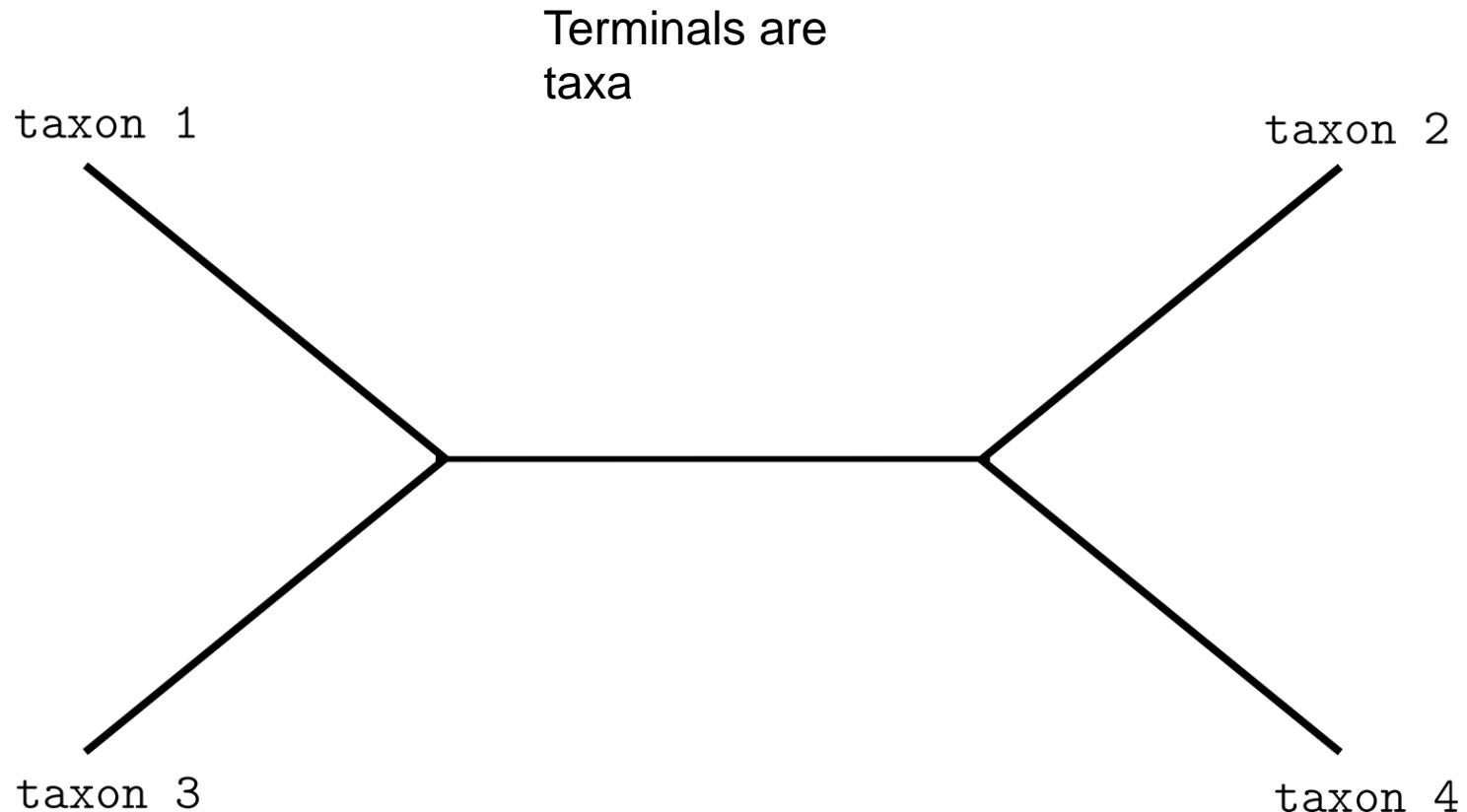


Fig. 1. The basal universal phylogenetic tree inferred from comparative analyses of rRNA sequences (4, 5). The root has been determined by using the paralogous gene couple EF-Tu/EFG (6).

Reading Trees



Reading Trees



Rooted / Unrooted Tree

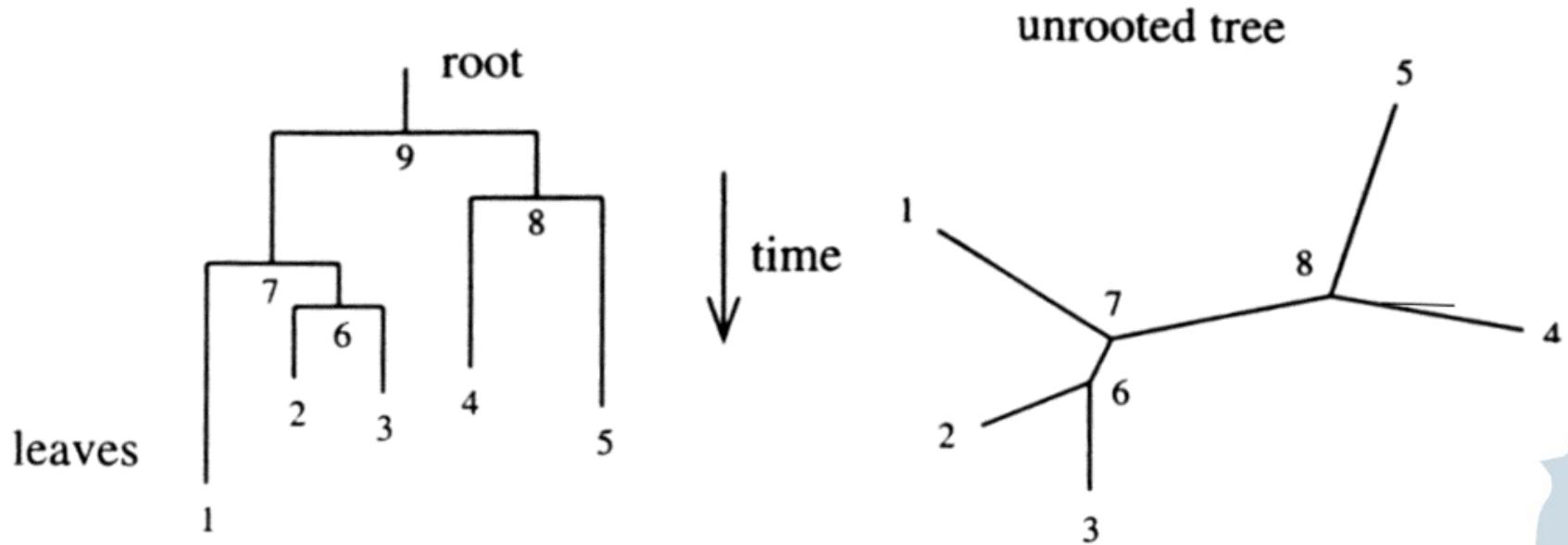
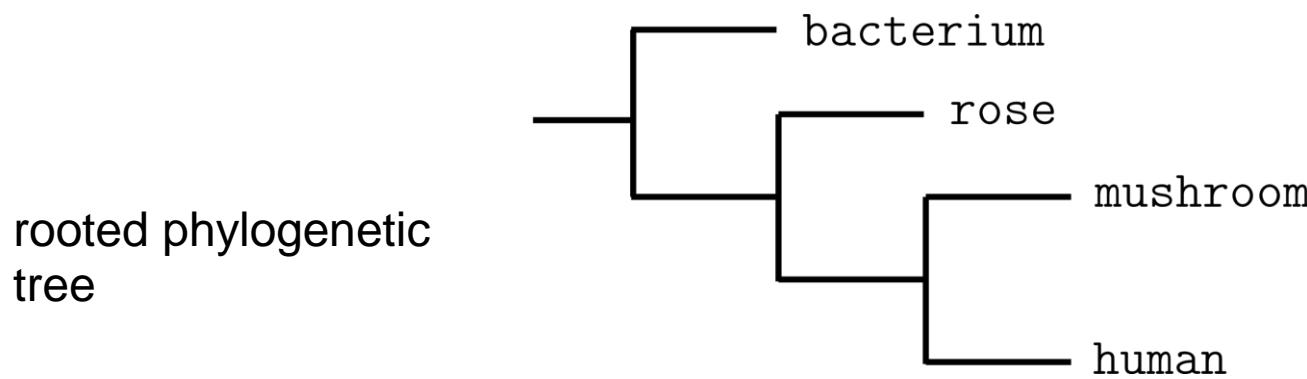
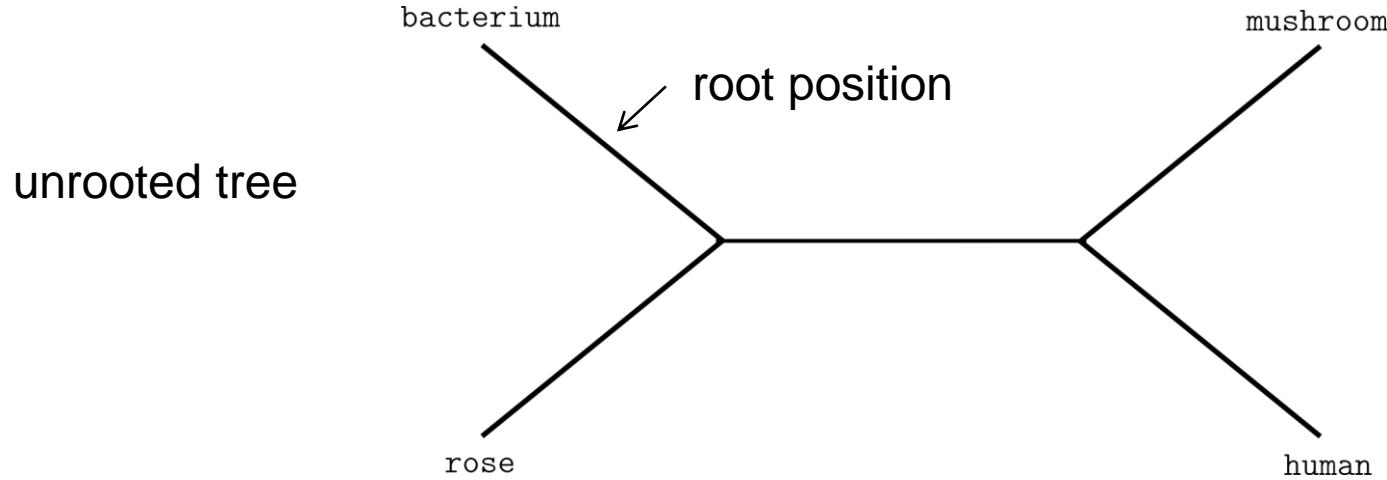
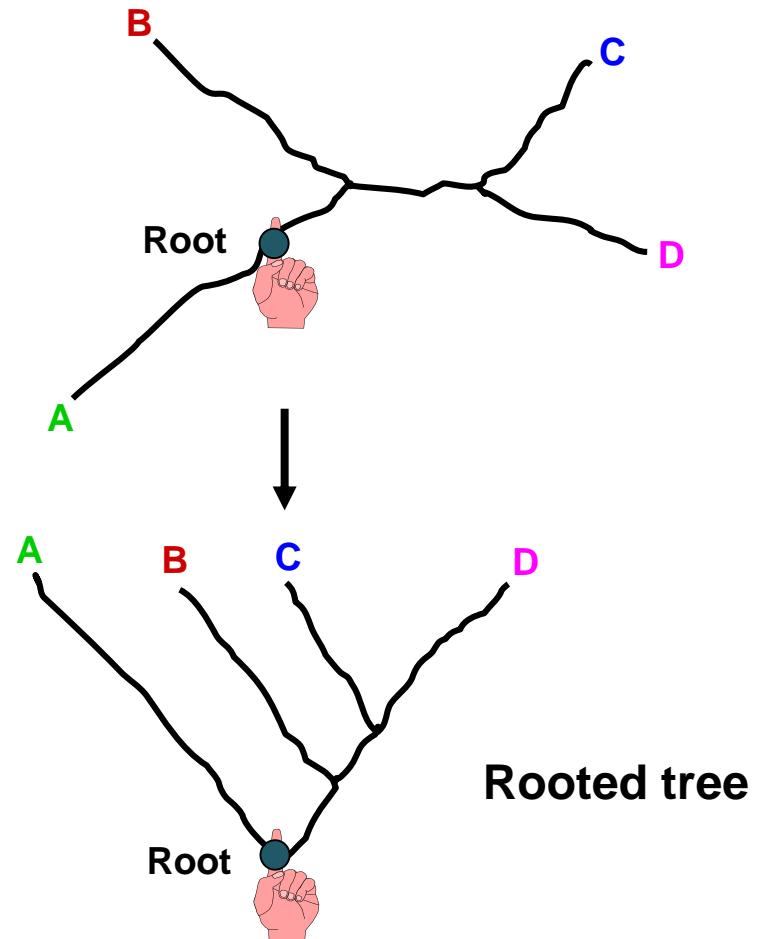


Figure 7.2 An example of a binary tree, showing the root and leaves, and the direction of evolutionary time (the most recent time being at the bottom of the figure). The corresponding unrooted tree is also shown; the direction of time here is undetermined.

Rooted / Unrooted Tree



Rooted / Unrooted Tree



Note that in this rooted tree, taxon A is no more closely related to taxon B than it is to C or D.

Counting Trees

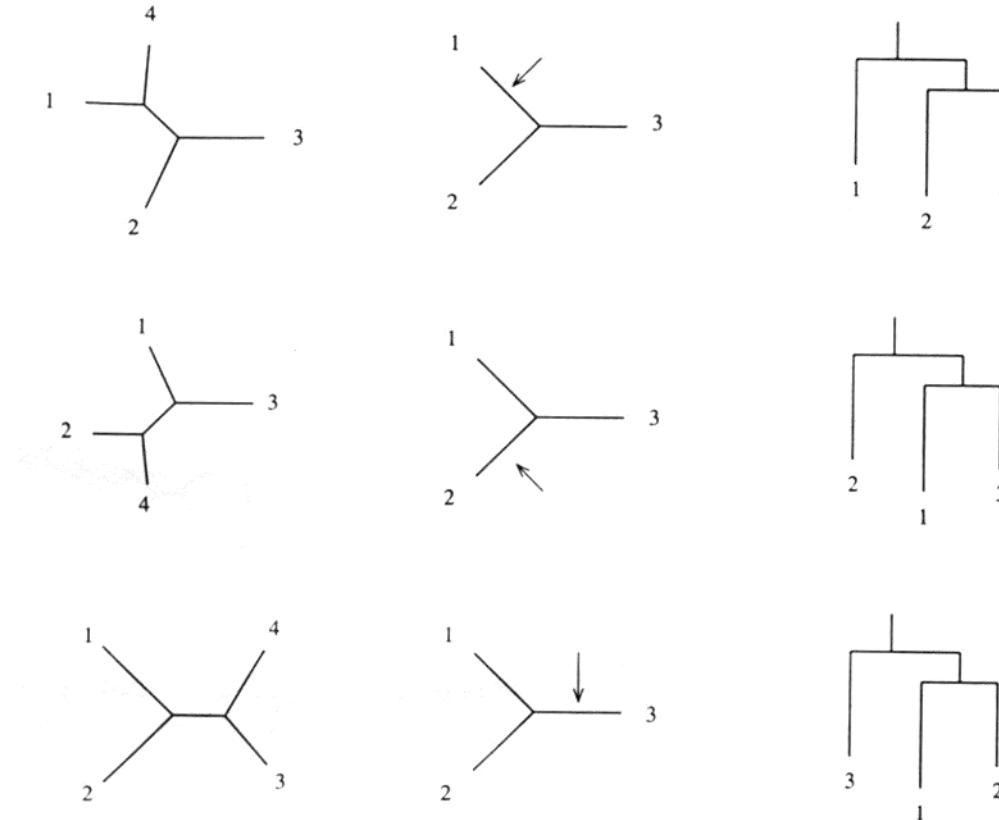
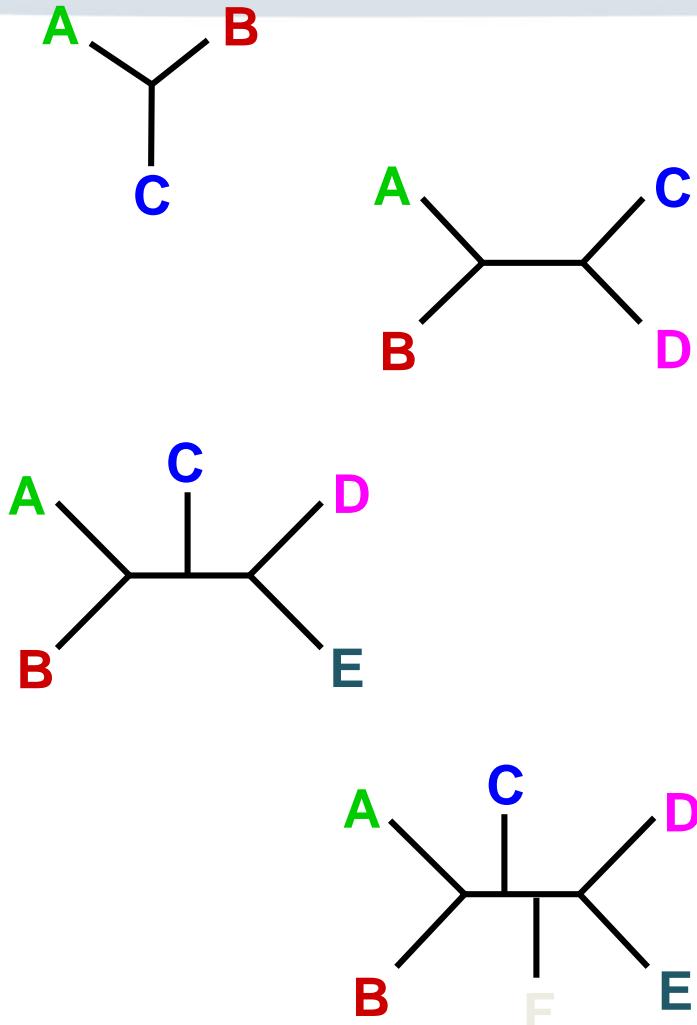


Figure 7.3 The rooted trees (right-hand column) derived from the unrooted tree for three sequences by picking different edges as positions for the root (arrows).

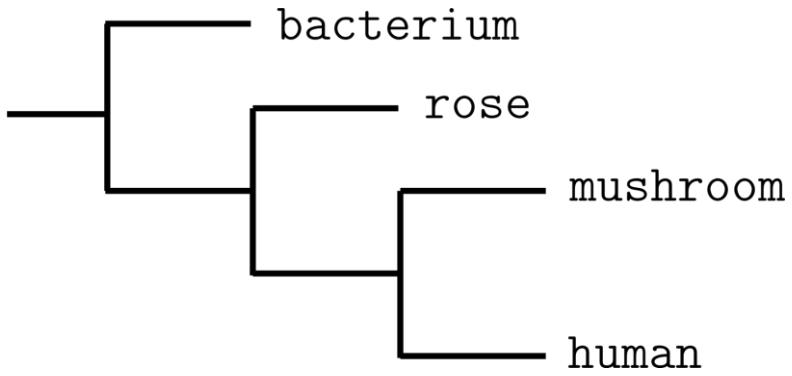
Counting Trees



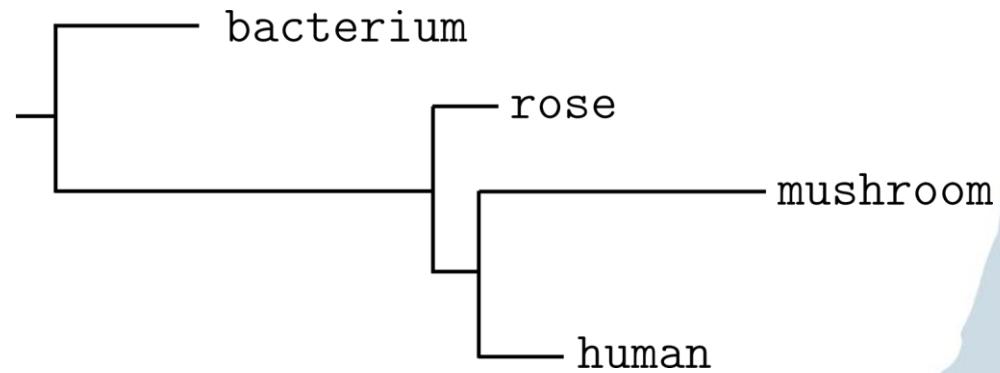
# Taxa (N)	# Unrooted trees
3	1
4	3
5	15
6	105
7	945
8	10,935
9	135,135
10	2,027,025
•	•
•	•
•	•
•	•
30	$\approx 3.58 \times 10^{36}$

$(2N - 5)!! = \# \text{ unrooted trees for } N \text{ taxa}$
 $(2N - 3)!! = \# \text{ rooted trees for } N \text{ taxa}$

Cladogram vs Phylogram

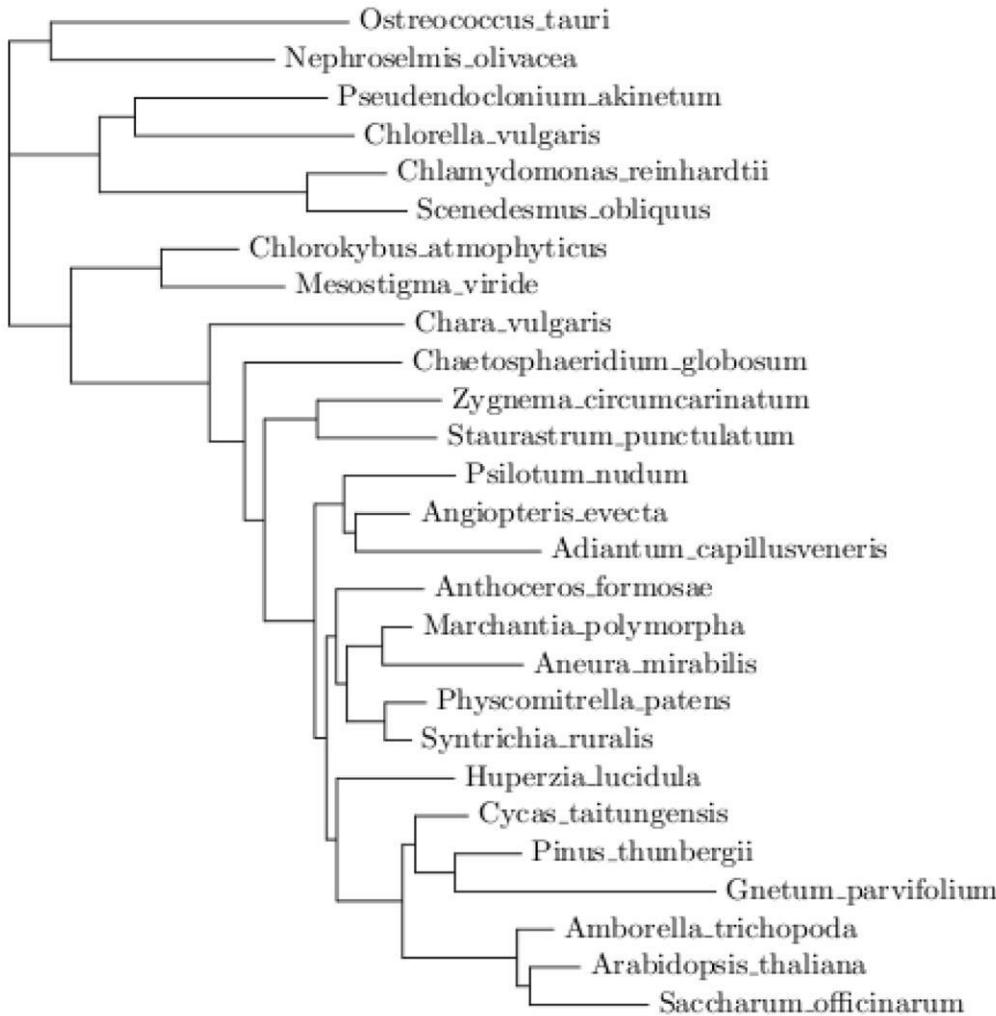


cladogram – arbitrary length branches



phylogram – branch length
proportional
to some measure of genetic distance

Outgroups



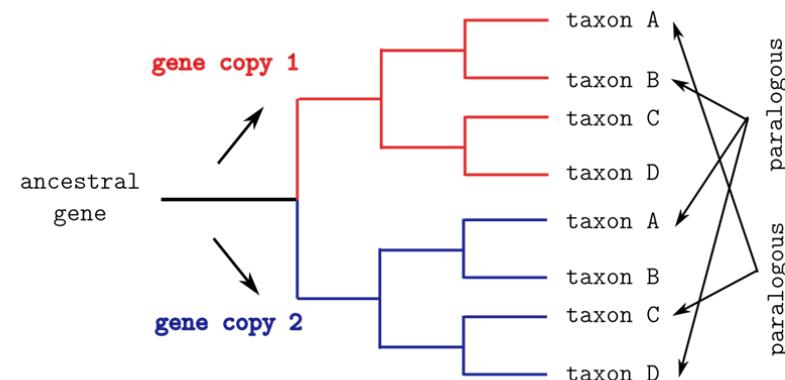
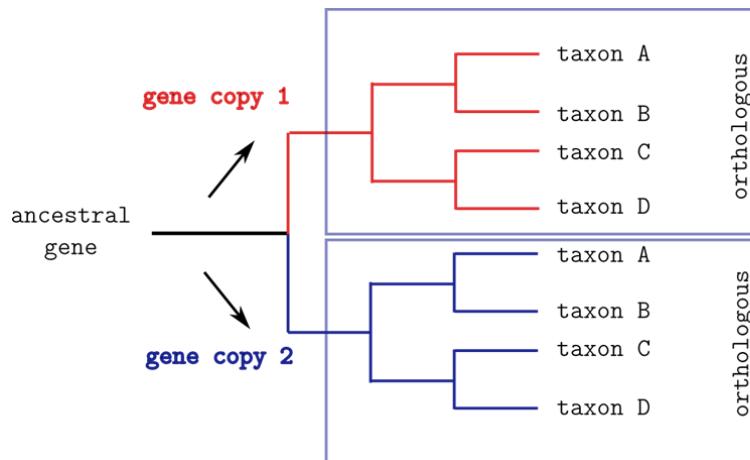
outgroup

ingroup

Orthologs and Paralogs

Orthology – homologous gene sequences

Paralogy – gene sequences separated by a gene duplication event



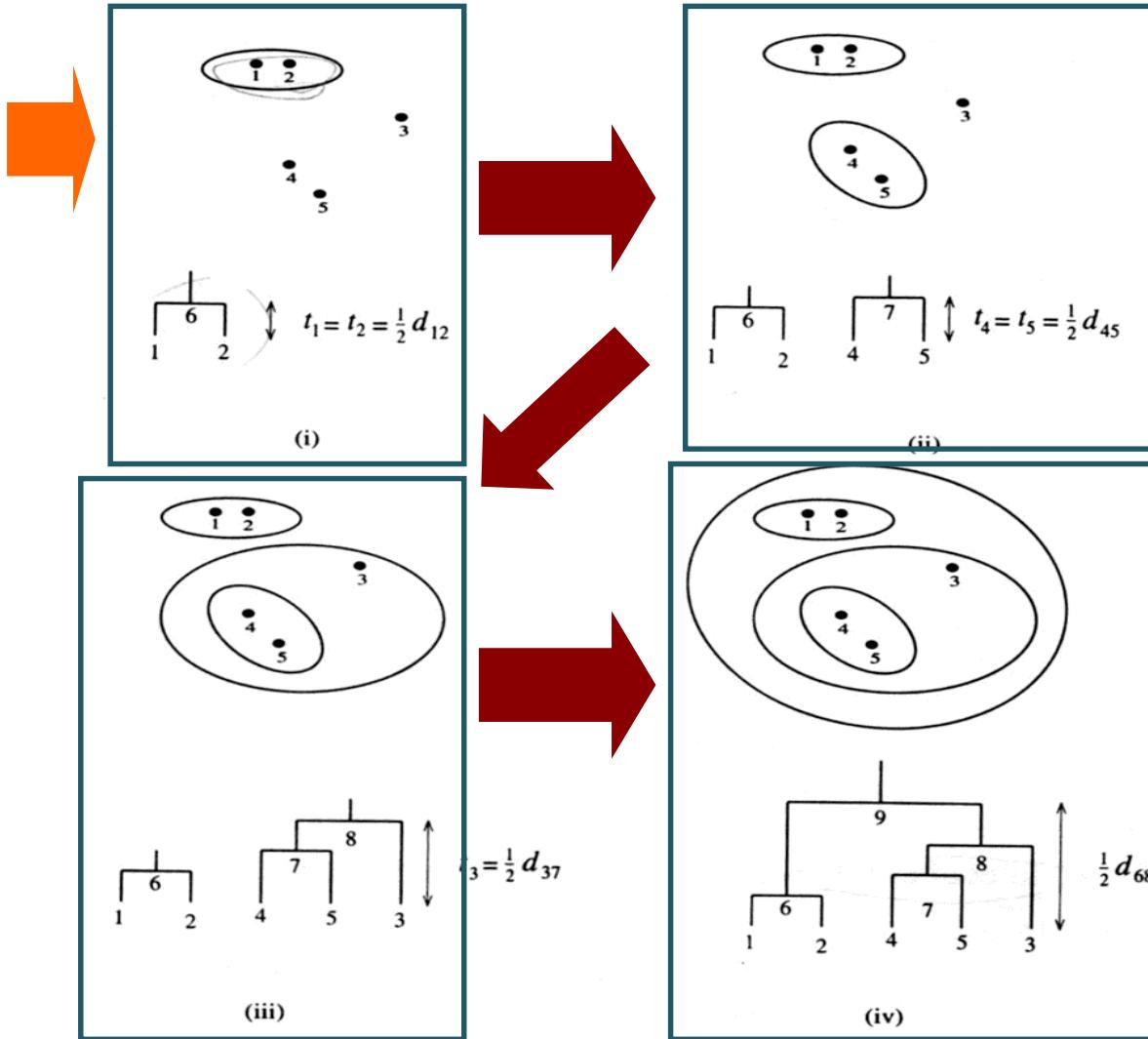
Rooting

- A phylogenetic analysis most often results in a formally unrooted network
- For phylogenetic analyses can be included an “outgroup” which will be used to root the tree
- The taxa of interest in are called the “ingroup”
- The assumption is that the ingroup taxa are more closely related to each other than any is to the outgroup
- If this assumption is wrong, then the interpretations of the phylogenetic tree will be wrong!

Methods

- Distance-based methods:
- Neighbor-joining
- UPGMA
- Character-based methods:
- Maximum parsimony
- Model-based methods:
 - Maximum likelihood
 - Bayesian inference

UPGMA



UPGMA

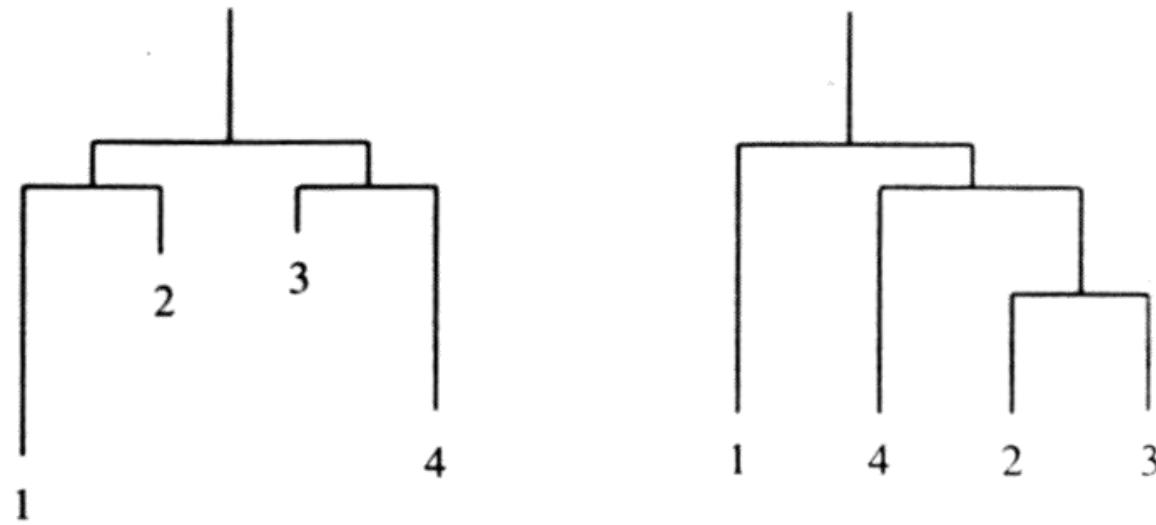
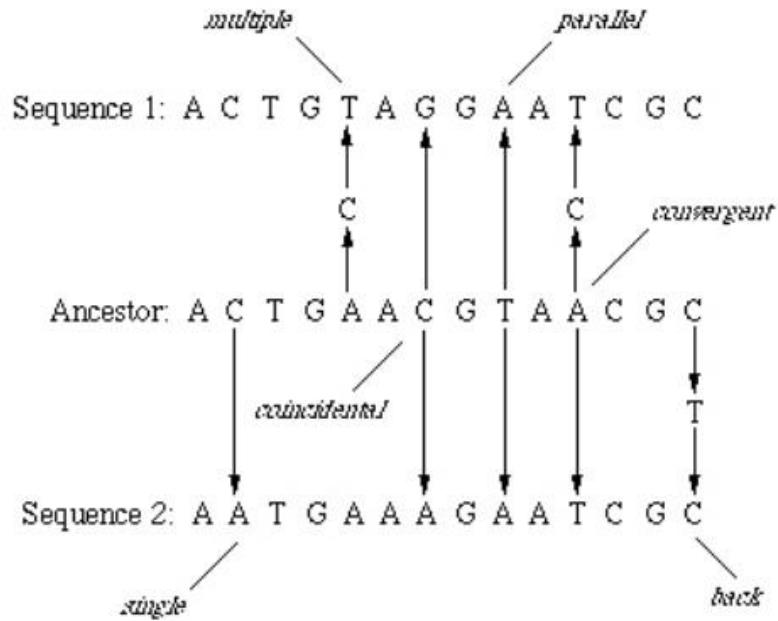


Figure 7.5 A tree (left) that is reconstructed incorrectly by UPGMA (right).

Neighbor Joining

- Very popular method
- Does not make molecular clock assumption : modified distance matrix constructed to adjust for differences in evolution rate of each taxon
- Produces unrooted tree
- Assumes additivity: distance between pairs of leaves = sum of lengths of edges connecting them

Neighbor Joining



Use models of substitution to correct these values

File Formats

- Newick (.nwk)

((species1:BranchLength,species2)Bootstrap,species 3);

- Advantages:
 - easy to perform
 - quick calculation
 - fit for sequences having high similarity scores
- Disadvantages:
 - the sequences are not considered as such (loss of information)
 - all sites are generally equally treated (do not take into account differences of substitution rates)
 - not applicable to distantly divergent sequences.

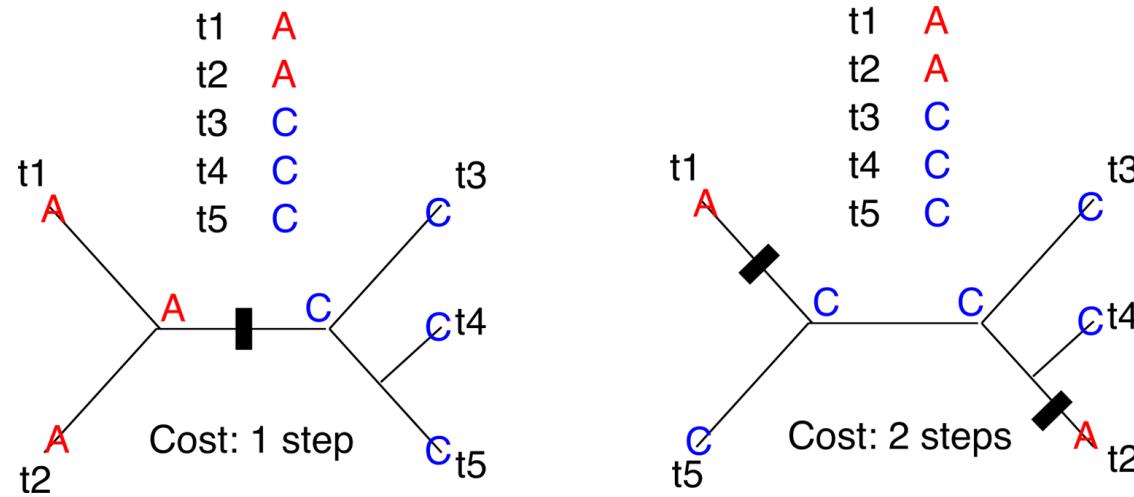
1. Requisitos Seaview e Mega
2. Descarregar sequencias na pasta hands_on_5
3. Alinhar usando Muscle em Aminoácidos
4. Fazer uma NJ de nucleótidos usando os parâmetros por omissão, 1000 bootstrap
5. Repetir o passo fazendo uma árvore de aminoácidos

Characters based tree

Maximum parsimony

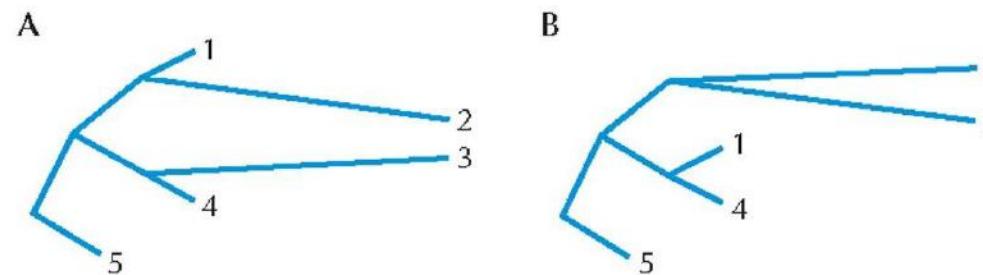
- Object is to **minimise the number of changes** necessary for the evolution of all characters on a tree.
- Character changes are *typically* treated as equally-weighted ie. the "cost" of changing from one state to another is the same between all states, but various weighting schemes can be applied
- Can be used with both morphological and molecular data, morphological characters may be ordered and polarised
- The tree with the **fewest changes/steps** is the MP tree. Might find many most- parsimonious solutions, which are often presented as a 50% majority-rule tree

Maximum parsimony



.Fitch's algorithm can be used to determine the most-parsimonious character reconstruction on any tree, the total score (or length) of a tree is the number of steps (changes) required by the most-parsimonious reconstructions of all characters, and the tree (or trees) with the lowest total score is the MP tree (or MP trees)

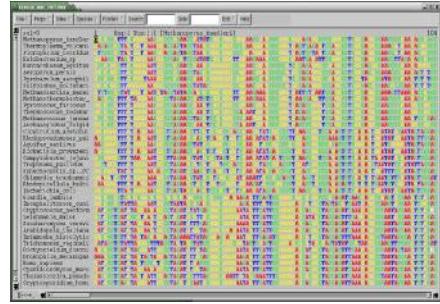
Long Branch Attraction



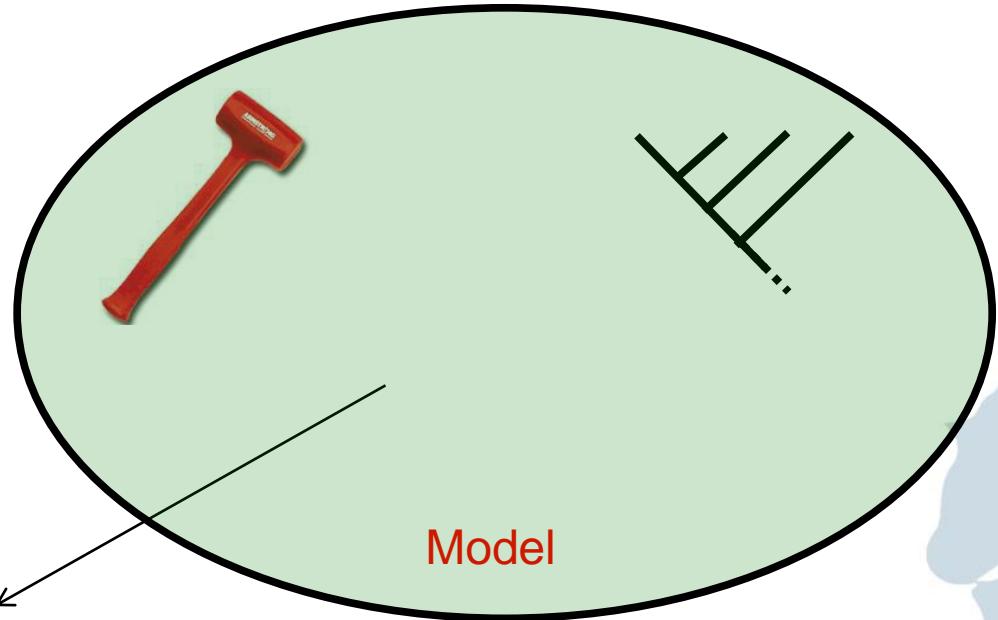
Maximum likelihood and Bayesian inference

- .Both use explicit models of character change that are evaluated on a tree using the likelihood function $\propto \text{Prob}(\mathcal{D} | \mathcal{H})$
- .They differ in their use of statistical paradigms

Model as mechanism of change and tree



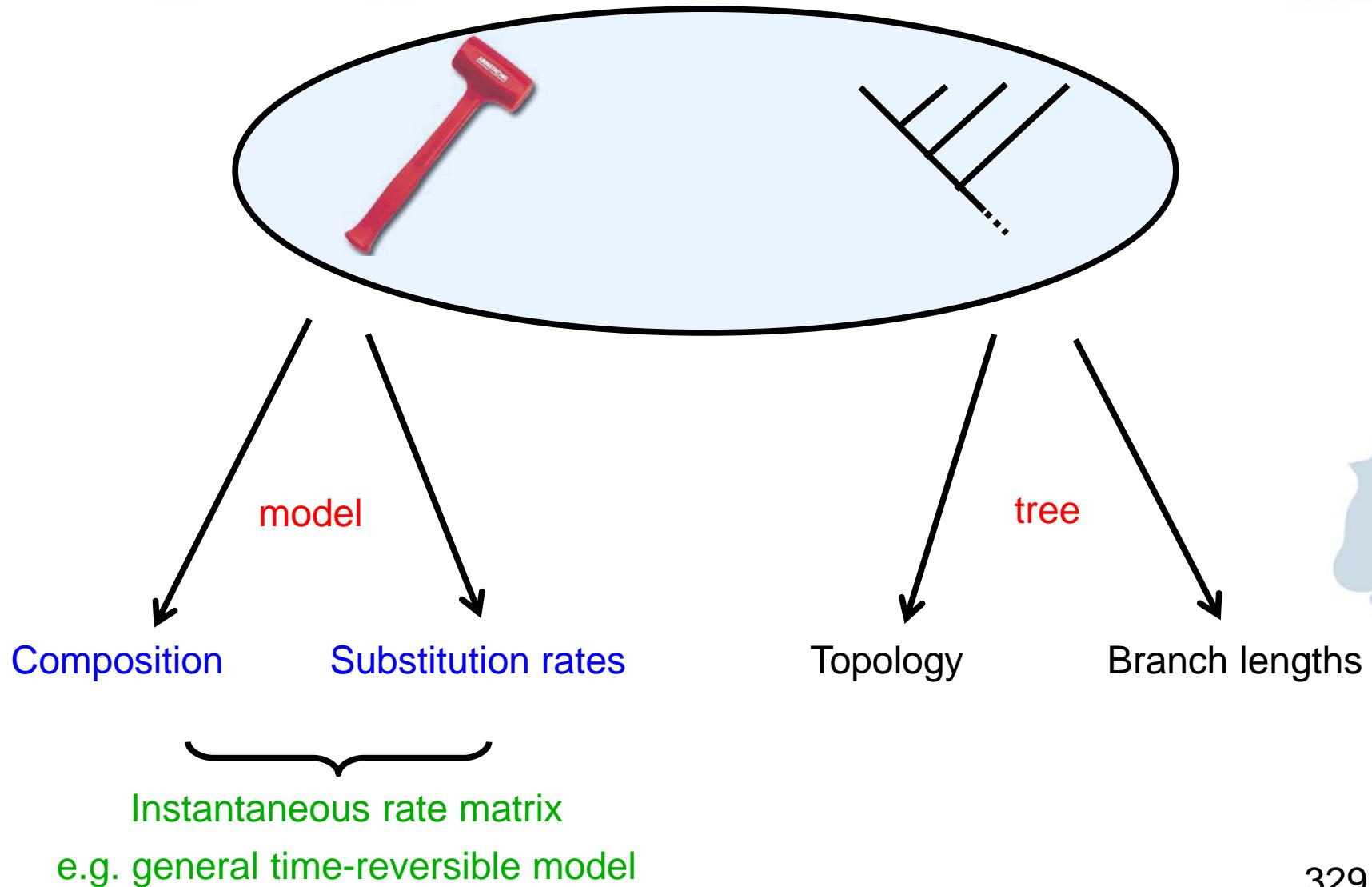
Data



$$\text{likelihood} \propto \text{Prob}(\mathcal{D} | \mathcal{H})$$

The likelihood is proportional to the probability of data given the hypothesis
(a model of character change plus tree topology)

Model as mechanism of change and tree



Models

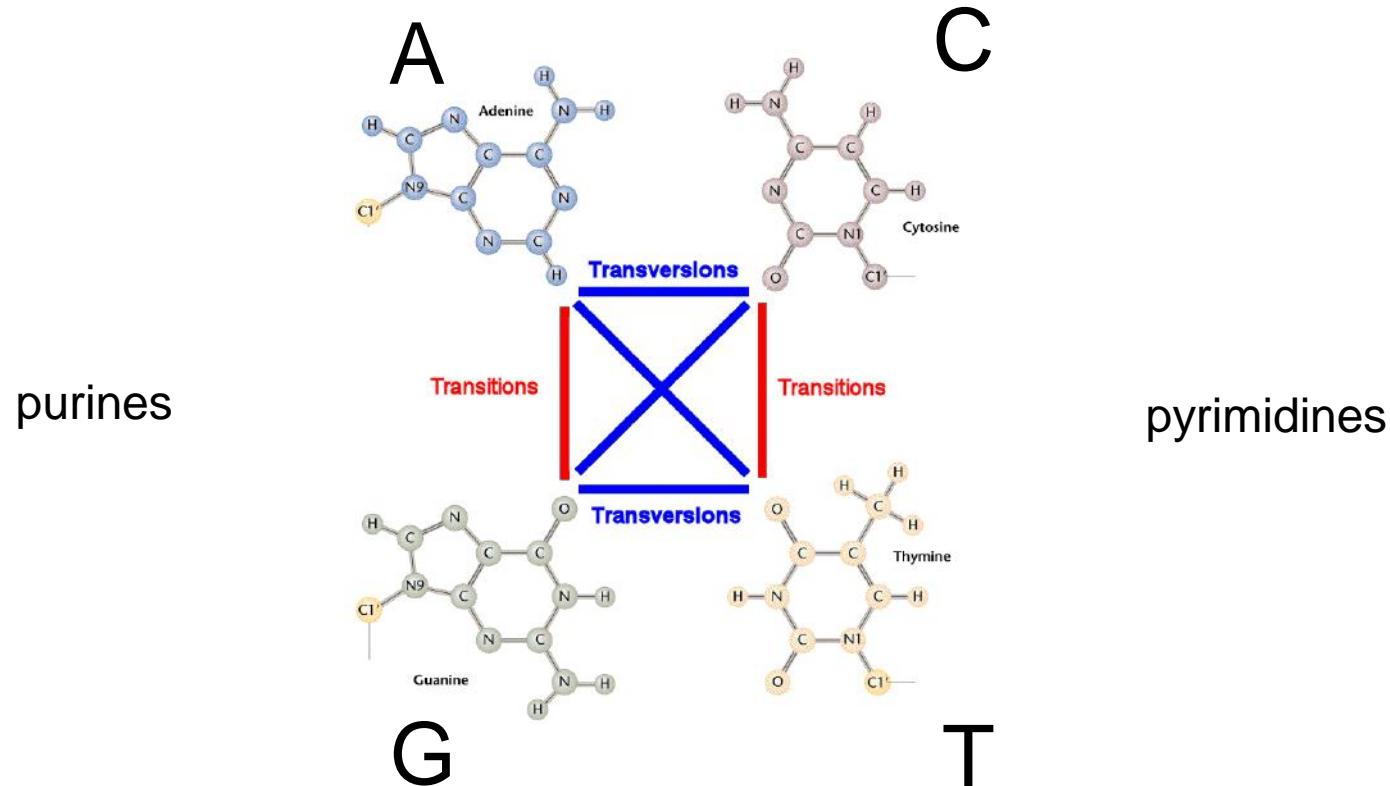
- 1) the base composition:
- π_a , π_c , π_g , & π_t
- The base composition frequency parameters remain constant over time (i.e. they are at equilibrium)
- They express the rate at which changes **to** each base occur
- Hence, the rate of change to a rare base would be low, whereas the change to a common base would be high
- » could be equal: 0.25, 0.25, 0.25, 0.25
- » or not: 0.3, 0.4, 0.2, 0.1
- » perhaps values are estimated from the data

Models

2) Substitution rates:

$$\mathcal{R} = \begin{pmatrix} & A & C & G & T \\ - & r_{A \rightarrow C} & r_{A \rightarrow G} & r_{A \rightarrow T} \\ r_{C \rightarrow A} & - & r_{C \rightarrow G} & r_{C \rightarrow T} \\ r_{G \rightarrow A} & r_{G \rightarrow C} & - & r_{G \rightarrow T} \\ r_{T \rightarrow A} & r_{T \rightarrow C} & r_{T \rightarrow G} & - \end{pmatrix} \begin{matrix} & A & C & G & T \\ A & & & & \\ C & & & & \\ G & & & & \\ T & & & & \end{matrix}$$

Transitions and transversions



Model GTR

$$Q = \begin{pmatrix} - & \mu r_{A \rightarrow C} \pi_C & \mu r_{A \rightarrow G} \pi_G & \mu r_{A \rightarrow T} \pi_T \\ \mu r_{C \rightarrow A} \pi_A & - & \mu r_{C \rightarrow G} \pi_G & \mu r_{C \rightarrow T} \pi_T \\ \mu r_{G \rightarrow A} \pi_A & \mu r_{G \rightarrow C} \pi_C & - & \mu r_{G \rightarrow T} \pi_T \\ \mu r_{T \rightarrow A} \pi_A & \mu r_{T \rightarrow C} \pi_C & \mu r_{T \rightarrow G} \pi_G & - \end{pmatrix}$$

$$GTR = \begin{matrix} & \textcolor{red}{A} & \textcolor{red}{C} & \textcolor{red}{G} & \textcolor{red}{T} \\ \textcolor{red}{A} & - & \mu r_i \pi_C & \mu r_j \pi_G & \mu r_k \pi_T \\ \textcolor{red}{C} & \mu r_i \pi_A & - & \mu r_l \pi_G & \mu r_m \pi_T \\ \textcolor{red}{G} & \mu r_j \pi_A & \mu r_l \pi_C & - & \mu r_n \pi_T \\ \textcolor{red}{T} & \mu r_k \pi_A & \mu r_m \pi_C & \mu r_n \pi_G & - \end{matrix}$$

GTR derived

Model derived from the *GTR* model – few have been implemented in phylogenetics

- **GTR** – unequal base frequencies and 6 substitution types
- **SYM** – equal base frequencies and 6 substitution types
- **HKY85** – unequal base frequencies and 2 substitution types (**transitions** and **transversion**)
- **F81** – unequal base frequencies and single substitution type
- **JC** – equal base frequencies and single substitution type

Model selection

2) The **Akaike Information Criterion**: $AIC_i = -2\log L_i + 2p_i$

where: i is the hypothesis (model + tree), and
 p is the number of free parameters

- Does not require models to be nested
- Calculate AIC for each model
- Choose model with lowest AIC score
- To be preferred over the LRT

Model selection

1) The **Likelihood Ratio Test** (LRT):

where: L_0 is a restricted (simpler) version of L_1

e.g.:

Null model = HKY+G

$-\ln L_0 = 7918.9556$

Alternative model = GTR+G

$-\ln L_1 = 7907.3330$

$2(\ln L_1 - \ln L_0) = 23.2451$

$df = 4$

P-value = 0.000113

Tools

- **Modeltest** conducts the LRT/AIC (and others) among a set of nested models
 - 14 substitution matrices with and without a *pinvar* and *gdasrv*
- Uses a crude distance tree to calculate the likelihoods of the models
- **MrModeltest** calculates similar for 24 models (those used by MrBayes)
- For amino acid models use **ProtTest** or **ModelGenerator**

Models parameters

- It is commonly recognised that not all sites evolve at the same rate – some may be constrained by selection. This can be incorporated into the model:
 1. *Site-specific rate categories* - defined *a priori*, e.g. first, second, third codon positions of a protein-coding gene.
 2. *Proportion of invariant sites* (pinvar) - assumes some proportion of sites is incapable of changing and all other sites vary at the same rate.
 3. *Gamma distributed among site rate variation* (gdasrv) - uses a number of discrete categories of rates that partitions a gamma distribution - the shape of the distribution is described by the parameter α

Amino acid substitution models

- **Poisson model** – (equiv. to JC) equal substitution rates and frequencies
- **Proportional model** – as Poisson but with unequal (empirically observed) frequencies

Empirically observed transition matrices:

- **Dayhoff** - derived from Dayhoff, et al.'s (1978) empirical substitution matrix
- **JTT** - Jones, Taylor, Thornton
- **WAG** - Whelan and Goldman

Bootstrap

- The bootstrap is a statistical method that is designed to test the reliability of the result by using **pseudo-replicates drawn from the original data**
- Draw characters/sites from the original data, with replacement, from the original data set to make a new one the same size. Repeat the phylogenetic analysis on this bootstrap replicate and repeat the process many times (100-1000).
- Interpretation of the bootstrap is difficult. It is known to be biased, but for a particular support values it not known where it is biased up or down. Its usually reckoned that 70% is statistically significant.

Bootstrap

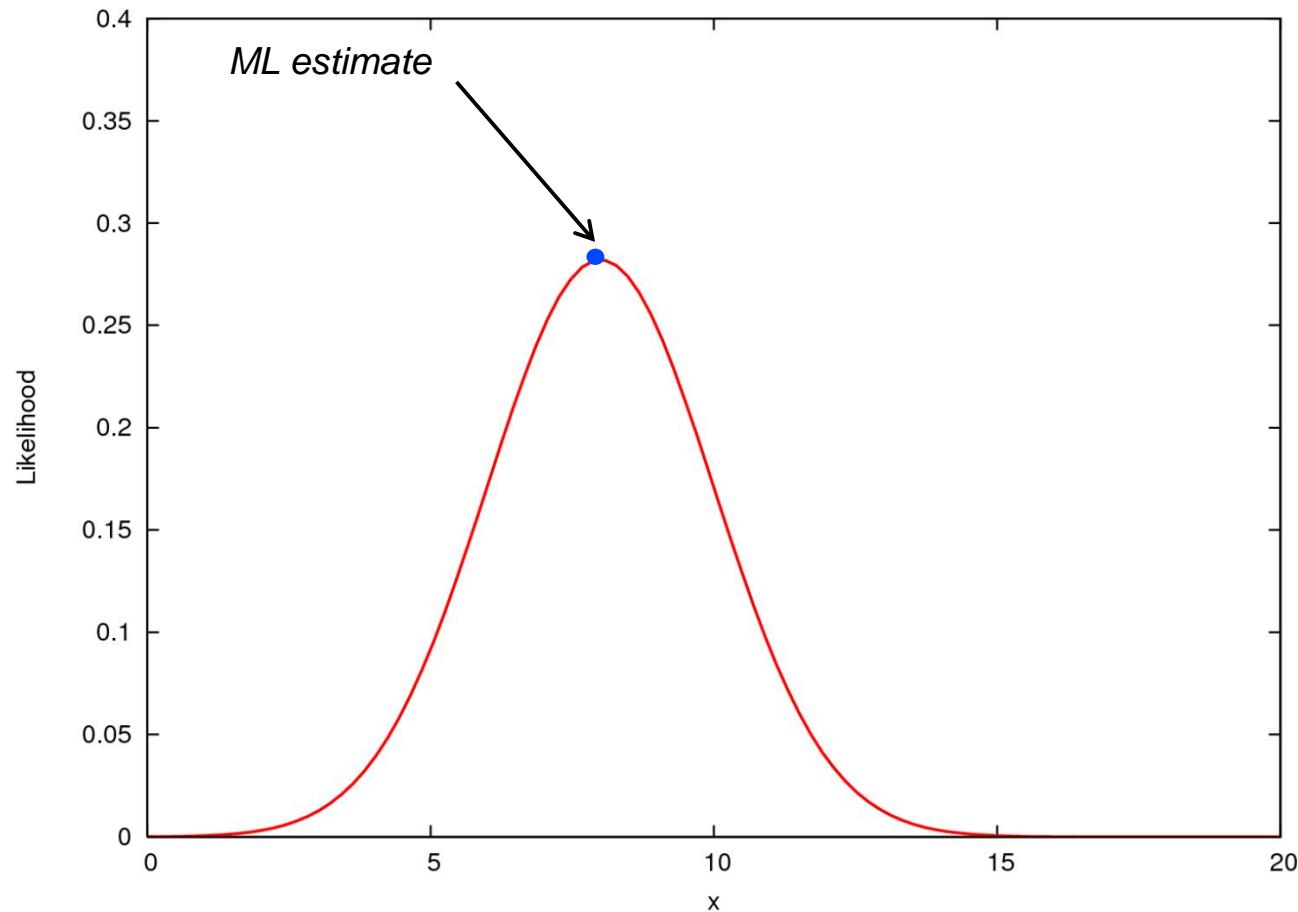
FGFR1_10 -----
 FGFR1_12 -----
 FGFR1_15 MWSWKCLLFWAFLVLTATLCTARPSPLPBC-----
 FGFR1_6 MWSWKCLLFWAFLVLTATLCTARPSPLPBC-----
 FGFR1_8 MWSWKCLLFWAFLVLTATLCTARPSPLPBC-----
 FGFR1_14 MWSWKCLLFWAFLVLTATLCTARPSPLPBCAQI PWGAPVEVESFLVHPGDLQLRCRIRDDVQSINWL RDGVQLAESNRTRITGEEVEVQDSVPADEGLYACVTSSPGSDDTTIFSVNVSDALPSSEDDDDDSSSEEKEIDNTKPN--P
 FGFR1_4 MWSWKCLLFWAFLVLTATLCTARPSPLPBCAQI PWGAPVEVESFLVHPGDLQLRCRIRDDVQSINWL RDGVQLAESNRTRITGEEVEVQDSVPADEGLYACVTSSPGSDDTTIFSVNVSDALPSSEDDDDDSSSEEKEIDNTKPN--P
 FGFR1_1 MWSWKCLLFWAFLVLTATLCTARPSPLPBCAQI PWGAPVEVESFLVHPGDLQLRCRIRDDVQSINWL RDGVQLAESNRTRITGEEVEVQDSVPADEGLYACVTSSPGSDDTTIFSVNVSDALPSSEDDDDDSSSEEKEIDNTKPN--P
 FGFR1_11 -----
 FGFR1_13 -----
 FGFR1_7 MWSWKCLLFWAFLVLTATLCTARPSPLPBC-----
 FGFR1_9 MWSWKCLLFWAFLVLTATLCTARPSPLPBC-----
 FGFR1_17 MWSWKCLLFWAFLVLTATLCTARPSPLPBC-----
 FGFR1_18 MWSWKCLLFWAFLVLTATLCTARPSPLPBC-----
 FGFR1_2 MWSWKCLLFWAFLVLTATLCTARPSPLPBCAQI PWGAPVEVESFLVHPGDLQLRCRIRDDVQSINWL RDGVQLAESNRTRITGEEVEVQDSVPADEGLYACVTSSPGSDDTTIFSVNVSDALPSSEDDDDDSSSEEKEIDNTKPN--P
 FGFR1_5 MWSWKCLLFWAFLVLTATLCTARPSPLPBCAQI PWGAPVEVESFLVHPGDLQLRCRIRDDVQSINWL RDGVQLAESNRTRITGEEVEVQDSVPADEGLYACVTSSPGSDDTTIFSVNVSDALPSSEDDDDDSSSEEKEIDNTKPN--P
 FGFR1_3 MWSWKCLLFWAFLVLTATLCTARPSPLPBCAQI PWGAPVEVESFLVHPGDLQLRCRIRDDVQSINWL RDGVQLAESNRTRITGEEVEVQDSVPADEGLYACVTSSPGSDDTTIFSVNVSDALPSSEDDDDDSSSEEKEIDNTKPN--P
 FGFR1_16 MWSWKCLLFWAFLVLTATLCTARPSPLPBCAQI PWGAPVEVESFLVHPGDLQLRCRIRDDVQSINWL RDGVQLAESNRTRITGEEVEVQDSVPADEGLYACVTSSPGSDDTTIFSVNVSDALPSSEDDDDDSSSEEKEIDNTKPN--P
 ruler 1.....10.....20.....30.....40.....50.....60.....70.....80.....90.....100.....110.....120.....130.....140.....150

FGFR1_10 -----MEKKLHAVPAAKTVIFKCPSSGTIPNPTLRLWLNKGKEFKPDHRIGGYKVRYATWSIIMDSVVPSDKGNYTCIVENEYGSINHTYQLDVVERSPHRFILQAGLPANKTVALGSNVFMCVKVYSPPOPHIQWLKHIEVNGSKI
 FGFR1_12 -----MEKKLHAVPAAKTVIFKCPSSGTIPNPTLRLWLNKGKEFKPDHRIGGYKVRYATWSIIMDSVVPSDKGNYTCIVENEYGSINHTYQLDVVERSPHRFILQAGLPANKTVALGSNVFMCVKVYSPPOPHIQWLKHIEVNGSKI
 FGFR1_15 VAPYWTSPPEKMEKKLHAVPAAKTVIFKCPSSGTIPNPTLRLWLNKGKEFKPDHRIGGYKVRYATWSIIMDSVVPSDKGNYTCIVENEYGSINHTYQLDVVERSPHRFILQAGLPANKTVALGSNVFMCVKVYSPPOPHIQWLKHIEVNGSKI
 FGFR1_6 VAPYWTSPPEKMEKKLHAVPAAKTVIFKCPSSGTIPNPTLRLWLNKGKEFKPDHRIGGYKVRYATWSIIMDSVVPSDKGNYTCIVENEYGSINHTYQLDVVERSPHRFILQAGLPANKTVALGSNVFMCVKVYSPPOPHIQWLKHIEVNGSKI
 FGFR1_8 VAPYWTSPPEKMEKKLHAVPAAKTVIFKCPSSGTIPNPTLRLWLNKGKEFKPDHRIGGYKVRYATWSIIMDSVVPSDKGNYTCIVENEYGSINHTYQLDVVERSPHRFILQAGLPANKTVALGSNVFMCVKVYSPPOPHIQWLKHIEVNGSKI
 FGFR1_14 VAPYWTSPPEKMEKKLHAVPAAKTVIFKCPSSGTIPNPTLRLWLNKGKEFKPDHRIGGYKVRYATWSIIMDSVVPSDKGNYTCIVENEYGSINHTYQLDVVERSPHRFILQAGLPANKTVALGSNVFMCVKVYSPPOPHIQWLKHIEVNGSKI
 FGFR1_4 VAPYWTSPPEKMEKKLHAVPAAKTVIFKCPSSGTIPNPTLRLWLNKGKEFKPDHRIGGYKVRYATWSIIMDSVVPSDKGNYTCIVENEYGSINHTYQLDVVERSPHRFILQAGLPANKTVALGSNVFMCVKVYSPPOPHIQWLKHIEVNGSKI
 FGFR1_1 VAPYWTSPPEKMEKKLHAVPAAKTVIFKCPSSGTIPNPTLRLWLNKGKEFKPDHRIGGYKVRYATWSIIMDSVVPSDKGNYTCIVENEYGSINHTYQLDVVERSPHRFILQAGLPANKTVALGSNVFMCVKVYSPPOPHIQWLKHIEVNGSKI
 FGFR1_11 -----MEKKLHAVPAAKTVIFKCPSSGTIPNPTLRLWLNKGKEFKPDHRIGGYKVRYATWSIIMDSVVPSDKGNYTCIVENEYGSINHTYQLDVVERSPHRFILQAGLPANKTVALGSNVFMCVKVYSPPOPHIQWLKHIEVNGSKI
 FGFR1_13 -----MEKKLHAVPAAKTVIFKCPSSGTIPNPTLRLWLNKGKEFKPDHRIGGYKVRYATWSIIMDSVVPSDKGNYTCIVENEYGSINHTYQLDVVERSPHRFILQAGLPANKTVALGSNVFMCVKVYSPPOPHIQWLKHIEVNGSKI
 FGFR1_7 VAPYWTSPPEKMEKKLHAVPAAKTVIFKCPSSGTIPNPTLRLWLNKGKEFKPDHRIGGYKVRYATWSIIMDSVVPSDKGNYTCIVENEYGSINHTYQLDVVERSPHRFILQAGLPANKTVALGSNVFMCVKVYSPPOPHIQWLKHIEVNGSKI
 FGFR1_9 VAPYWTSPPEKMEKKLHAVPAAKTVIFKCPSSGTIPNPTLRLWLNKGKEFKPDHRIGGYKVRYATWSIIMDSVVPSDKGNYTCIVENEYGSINHTYQLDVVERSPHRFILQAGLPANKTVALGSNVFMCVKVYSPPOPHIQWLKHIEVNGSKI
 FGFR1_17 VAPYWTSPPEKMEKKLHAVPAAKTVIFKCPSSGTIPNPTLRLWLNKGKEFKPDHRIGGYKVRYATWSIIMDSVVPSDKGNYTCIVENEYGSINHTYQLDVVERSPHRFILQAGLPANKTVALGSNVFMCVKVYSPPOPHIQWLKHIEVNGSKI
 FGFR1_18 VAPYWTSPPEKMEKKLHAVPAAKTVIFKCPSSGTIPNPTLRLWLNKGKEFKPDHRIGGYKVRYATWSIIMDSVVPSDKGNYTCIVENEYGSINHTYQLDVVERSPHRFILQAGLPANKTVALGSNVFMCVKVYSPPOPHIQWLKHIEVNGSKI
 FGFR1_2 VAPYWTSPPEKMEKKLHAVPAAKTVIFKCPSSGTIPNPTLRLWLNKGKEFKPDHRIGGYKVRYATWSIIMDSVVPSDKGNYTCIVENEYGSINHTYQLDVVERSPHRFILQAGLPANKTVALGSNVFMCVKVYSPPOPHIQWLKHIEVNGSKI
 FGFR1_5 VAPYWTSPPEKMEKKLHAVPAAKTVIFKCPSSGTIPNPTLRLWLNKGKEFKPDHRIGGYKVRYATWSIIMDSVVPSDKGNYTCIVENEYGSINHTYQLDVVERSPHRFILQAGLPANKTVALGSNVFMCVKVYSPPOPHIQWLKHIEVNGSKI
 FGFR1_3 -----CPDLEBAKSCSASFHSI
 FGFR1_16 -----PWGAPVEVBSELVHPGDLQLRCRIRDDVQSINWL RDGVQLAESNRTRITGEEVEVQDSVPADSGLYACVTSSPGSDDTTIFSVNVSDALPSSEDDDDDSSSEEKEIDNTKPN--P
 ruler160.....170.....180.....190.....200.....210.....220.....230.....240.....250.....260.....270.....280.....290.....300

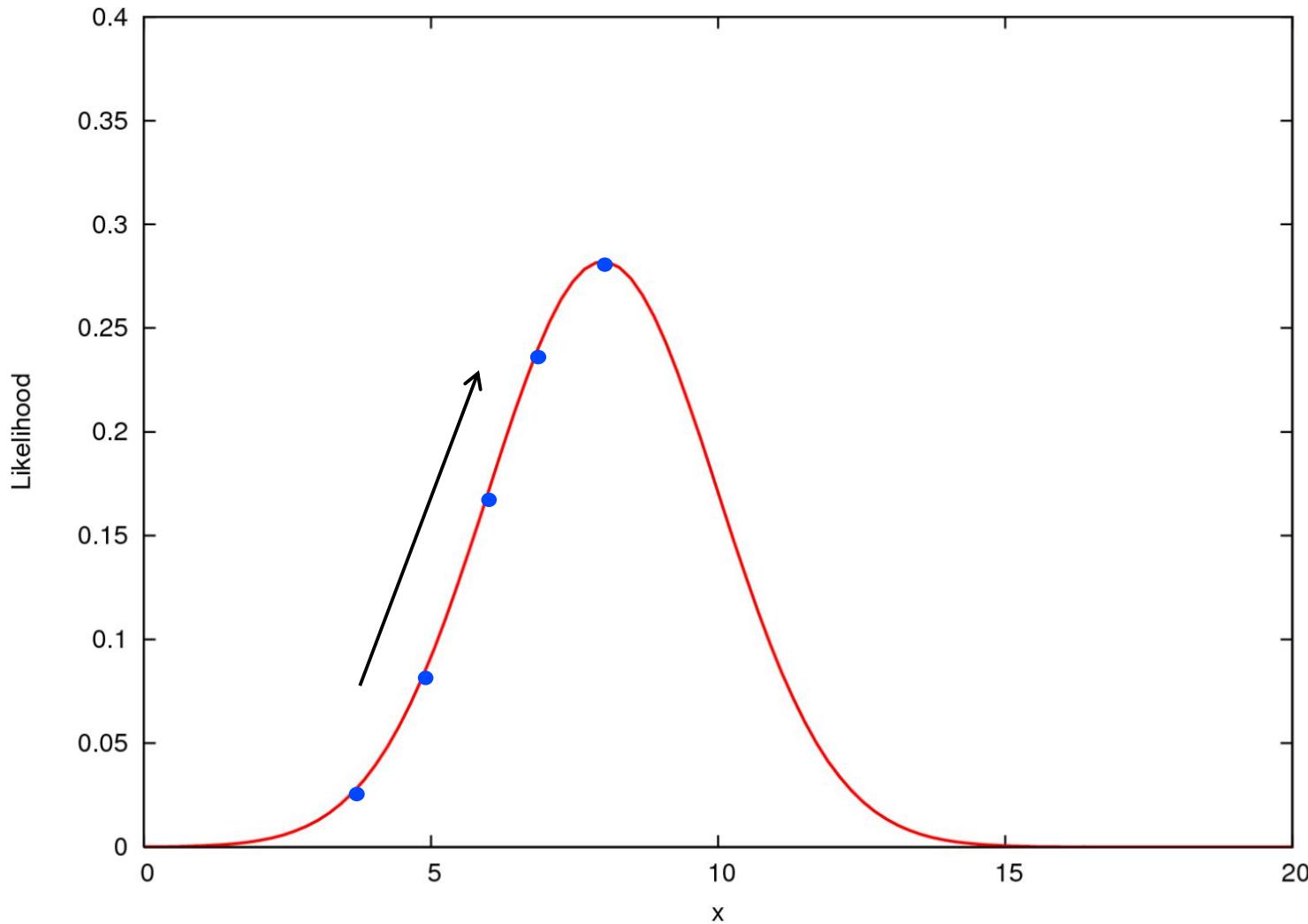
Hands On 6

1. Requisitos Seaview+jModelTest e/ou Mega
2. Alinhar usando MUSCLE
3. Determinar modelo evolucionário
4. Fazer uma ML usando 100 de bootstrap e o modelo determinado no passo 3

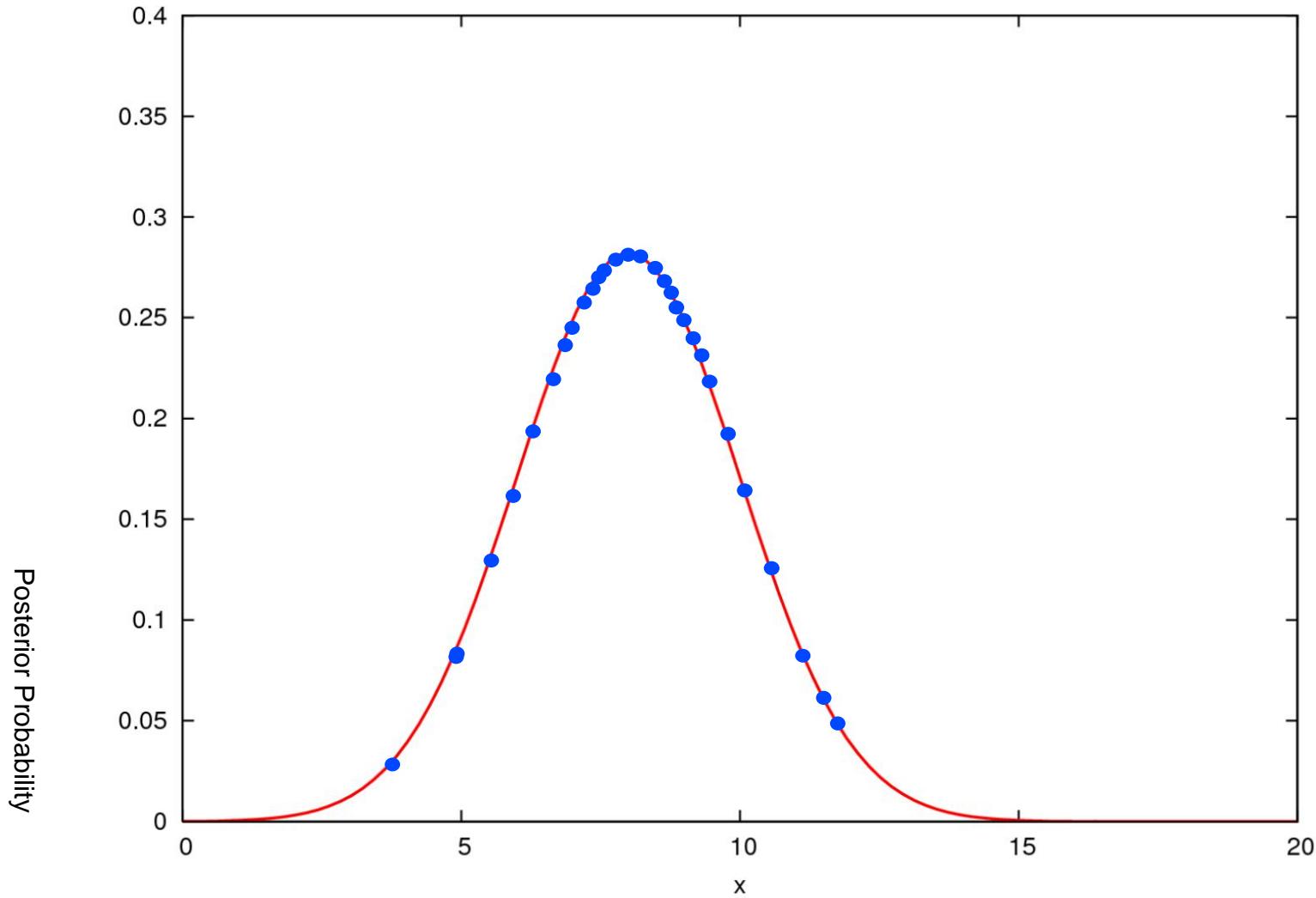
ML and PP density



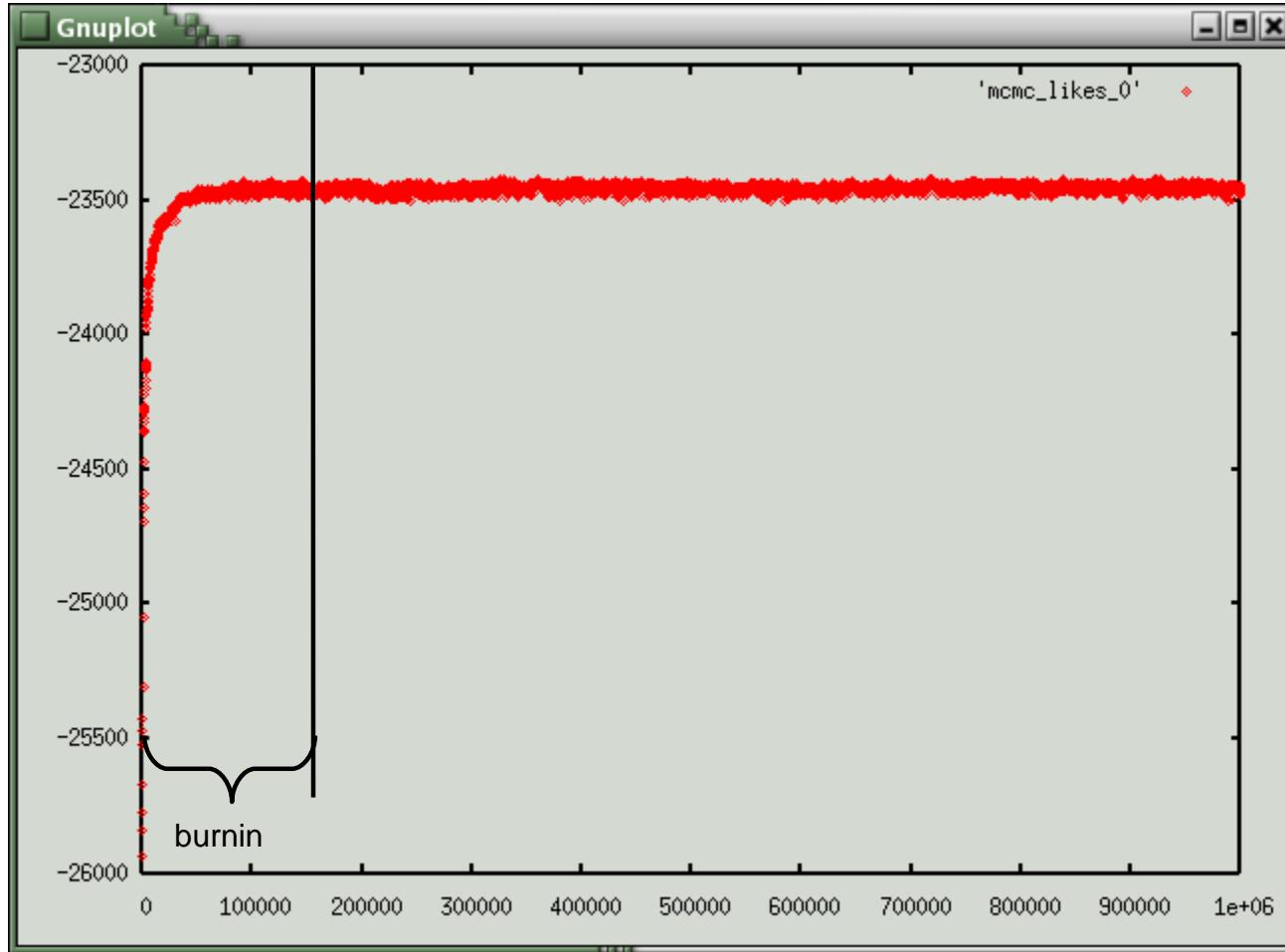
ML and PP density



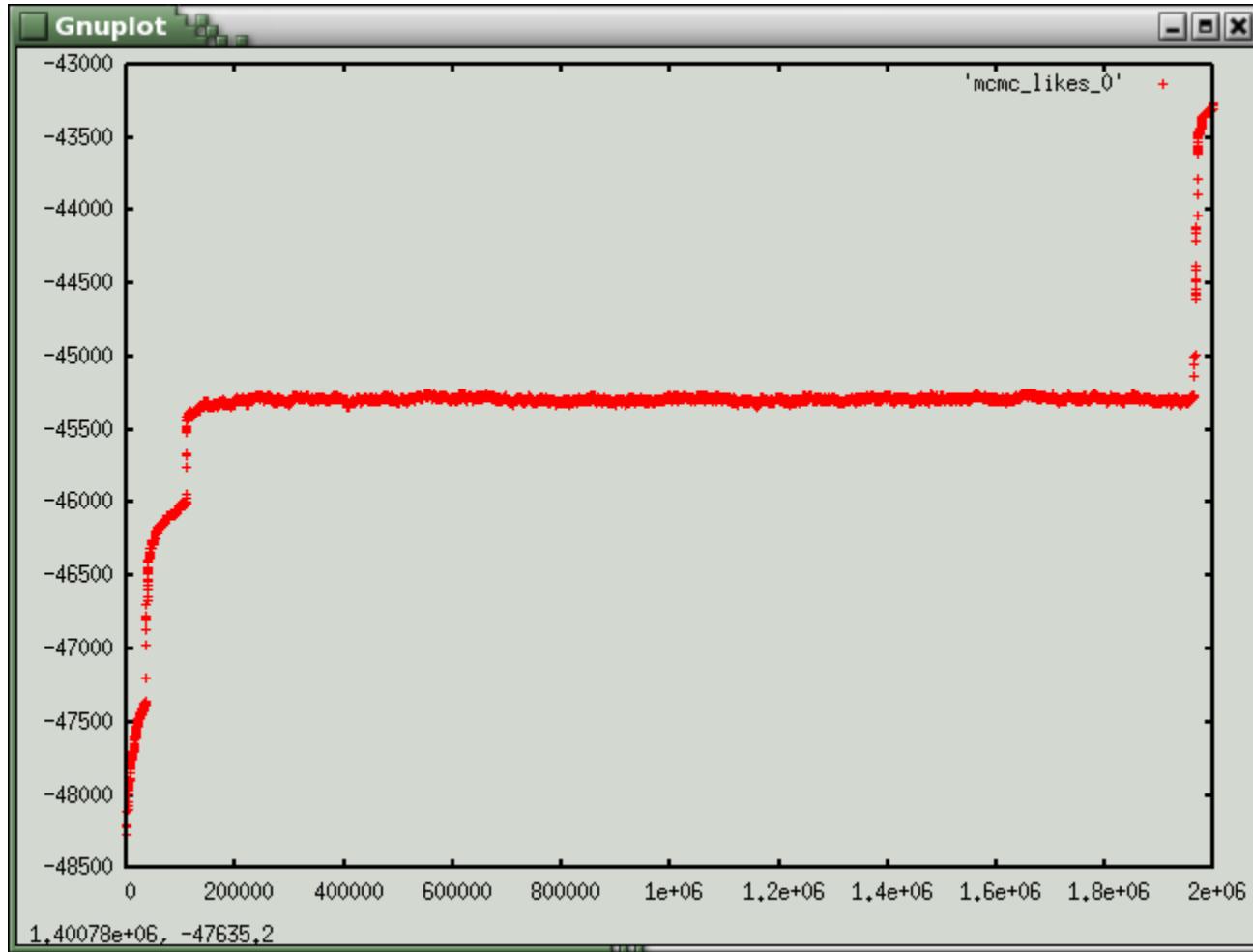
ML and PP density



Likelihood plot



Likelihood plot



Mcmc output

gens	Composition				Substitution Rate					gdasrv	
	A	C	G	T	A-C	A-G	A-T	C-G	C-T	G-T	
→ 100	0.22515466	0.24584661	0.30733981	0.22165891	0.11970749	0.23917488	0.08580185	0.14260800	0.32394051	0.08876728	0.874116
200	0.22581011	0.23935950	0.28160072	0.25322967	0.11361513	0.22751384	0.10471647	0.16653072	0.26859341	0.11903044	1.930917
300	0.23172551	0.27706732	0.27019486	0.22101230	0.12093299	0.20611349	0.08865832	0.15770240	0.28720945	0.13938334	4.380087
400	0.22999880	0.26063025	0.28697633	0.22239461	0.09333566	0.24749557	0.12001019	0.14071384	0.29772513	0.10071961	10.893358
500	0.24616074	0.25080719	0.27680877	0.22622330	0.10333398	0.20508574	0.10527309	0.16562894	0.31464847	0.10602977	14.875529
600	0.21219225	0.28281963	0.28884274	0.21614538	0.11325671	0.25977835	0.12678584	0.11935713	0.28575591	0.09506605	8.521777
etc...											
→ tree t_100 = [&U] (((1:0.263151, 2:0.0564195):0.206267, 3:0.223034):0.73243, 4:0.107335, 5:0.0742962); tree t_200 = [&U] (((5:0.115119, 4:0.0936513):1.14124, 2:0.0686334):0.206054, 3:0.338046, 1:0.223061); tree t_300 = [&U] ((2:0.0799976, (4:0.0892171, 5:0.115119):1.10016):0.174534, 3:0.461212, 1:0.236187); tree t_400 = [&U] ((2:0.0800969, (5:0.126667, 4:0.09186):1.39515):0.21926, 3:0.445261, 1:0.335234); tree t_500 = [&U] (((4:0.100586, 5:0.149979):1.92335, 2:0.0800969):0.268639, 1:0.335234, 3:0.556183); tree t_600 = [&U] (((3:0.671303, 1:0.413087):0.347891, 2:0.178729):2.18893, 5:0.0845881, 4:0.143907); etc...											

- Note that at each generation the parameter values are known (i.e. they are the current values of the chain) hence the likelihood is easy and quick to calculate, this makes BA a relatively quick method when compared to ML

Likelihood plot

- Are we there yet?

AWTY is a system for the graphical exploration of MCMC convergence, written by Jim Wilgenbusch, Dan Warren, and David Swofford.

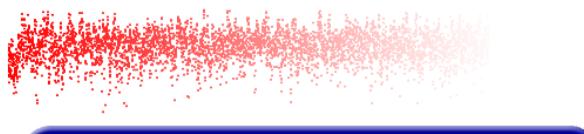
AWTY online

[About AWTY](#)

[Start a new session](#)

[Return to an old session](#)

[Contact us](#)



About AWTY

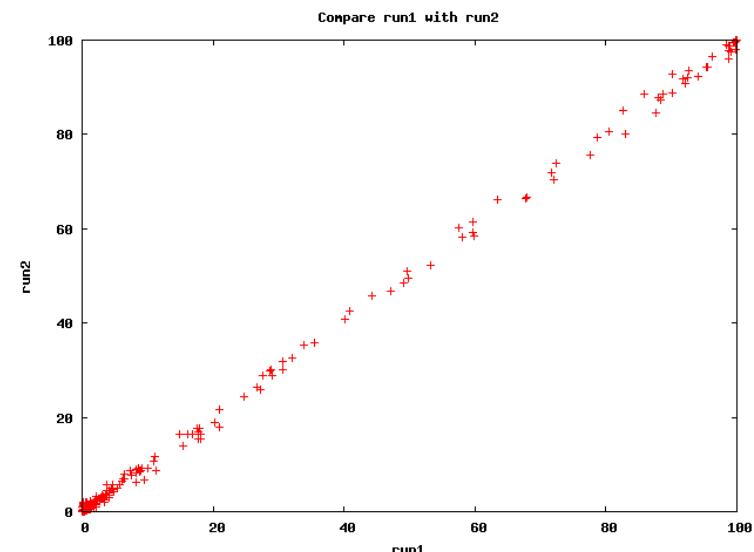
AWTY is a system for graphical exploration of Markov chain Monte Carlo (MCMC) convergence in Bayesian phylogenetic inference. The graphics produced by AWTY are designed to help assess whether an MCMC analysis has run long enough, such that tree topologies are being sampled in proportion to their true posterior probability distribution. In other words, "Are We There Yet?" or AWTY for short. Admittedly, the results generated by AWTY will never be able to answer this question with a definitive yes; however, in some cases results will point confidently to the answer no. See the [AWTY image gallery](#) for some examples.

To produce plots in AWTY a NEXUS or NEWICK formatted tree file representing a set of trees sampled over an MCMC run is required. To date, tree files generated by [MrBayes](#) and [BAMBE](#) have been tested. AWTY provides several graphical formats to display results or results may also be downloaded and analyzed using the plotting package of your choice.

The online version of AWTY is written in [Perl](#) and [PHP](#). Posterior probabilities of splits and topological tree distances are calculated by [PAUP*](#). Graphics are generated by [Gnuplot](#).

Citation:

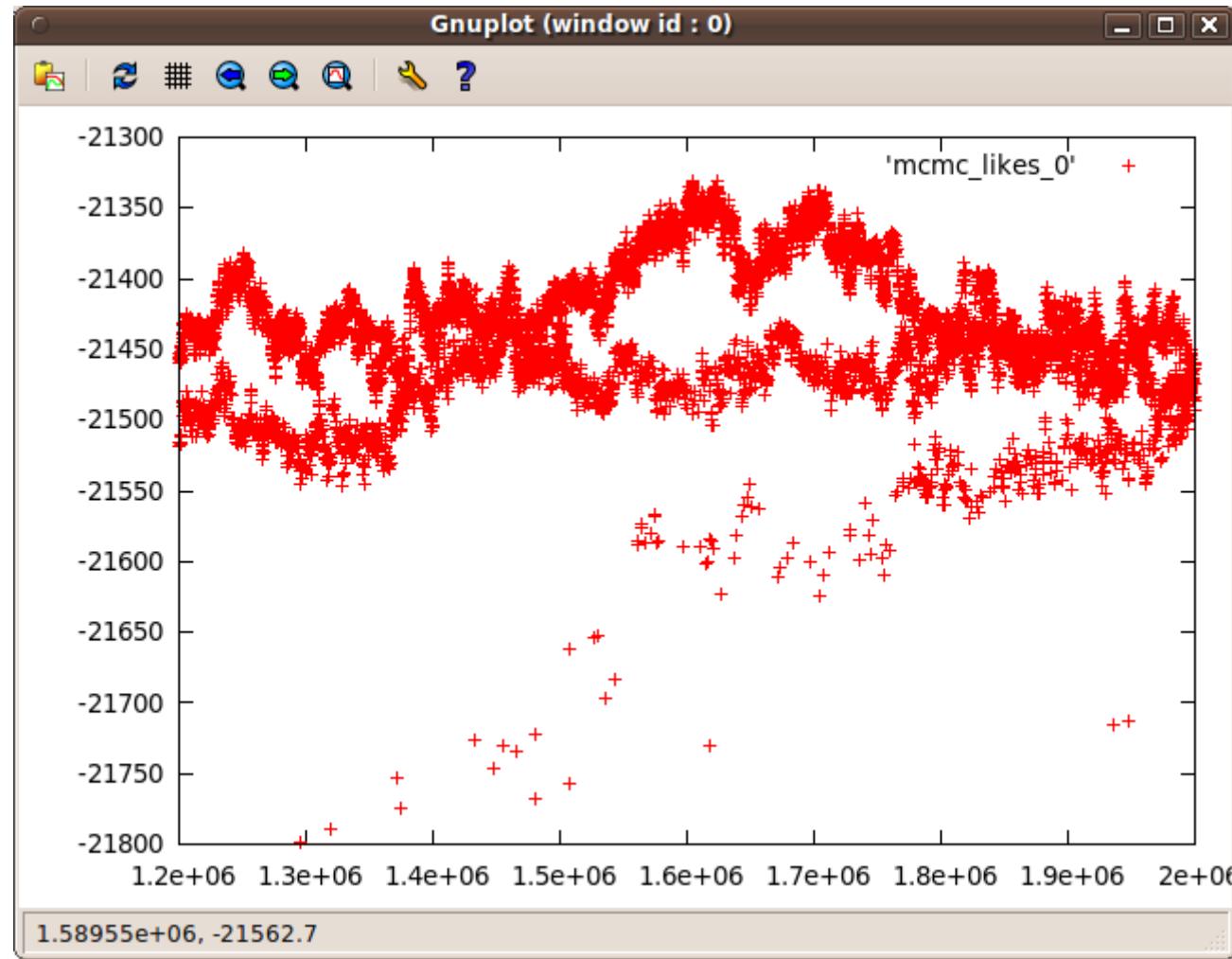
Wilgenbusch J.C., Warren D.L., Swofford D.L. 2004. AWTY: A system for graphical exploration of MCMC convergence in Bayesian phylogenetic inference.
<http://ceb.csit.fsu.edu/awty>.



Convergence

- Run the analysis more than once and check that the separate runs give similar results
- Monitor the average standard deviation of split support between two separate runs (MrBayes does this by default)

Likelihood plot

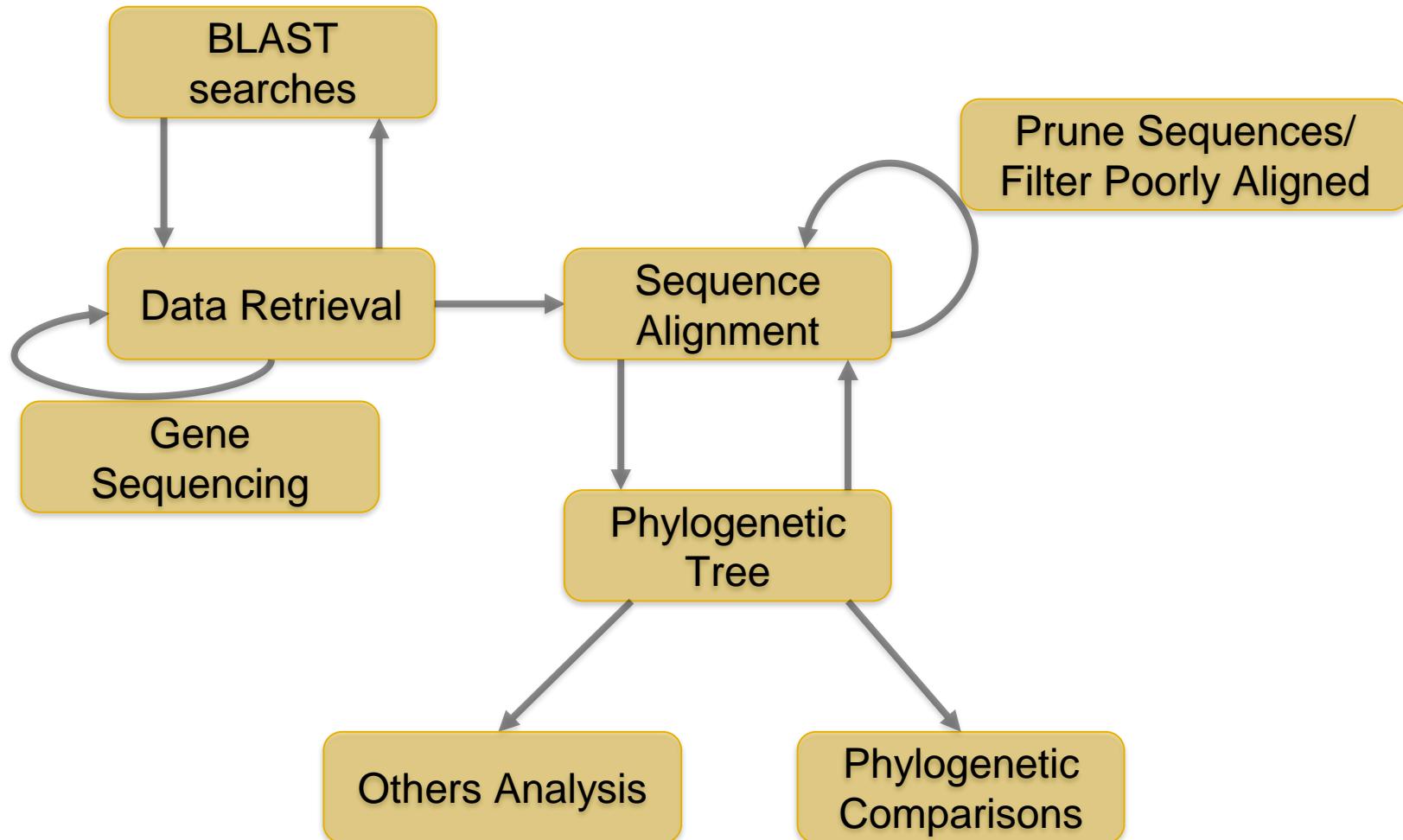


Alternatives

- <http://www.thines-lab.senckenberg.de/simba/>
- <http://phylemon.bioinfo.cipf.es/>
- http://phylogeny.lirmm.fr/phylo_cgi/index.cgi

1. Requisitos Mr. Bayes e TreeGraph
2. Abrir executável do Mr.Bayes (32 ou 64 bits)
3. “Execute nomedoficheiro.nex”
4. “help lset”
5. “lset nst=6 rates=invgamma” (modelo GTR+I+G)
6. “help mcmc”
7. “mcmc ngen=200 000”
8. “sump” (confirmar parâmetros)
9. “sumt conformat=simple”
10. Executar TreeGraph e abrir ficheiro .con gerado no Mr. Bayes

Typical Workflow



Hands On 8

1. Requisitos Tree-Puzzle
2. Abrir puzzle-windows-mingw.exe
3. k + enter -> K Tree search procedure? Evaluate user defined trees
4. m + enter até ao modelo GTR
5. w + enter até -> w Model of rate heterogeneity? Mixed (1 invariable + 4 Gamma rates)
6. y + enter
7. my_trees.txt (ficheiros com as árvores)
8. Abrir ficheiro .puzzle
9. No final do ficheiro aparece a comparação entre as 3 árvores “COMPARISON OF USER TREES (NO CLOCK)”

Data Warehouses Again

EMBL-EBI Treefam

Services | Research | Training | About us

Search Examples: BRCA2, ENSP0000428982, or do a sequence search

Home | Search | Browse | Download | Help | Forum

Family: awaiting annotation (TF337278)
Description: awaiting annotation

Summary

- Gene Tree
- Wikipedia
- Sequences
- Downloads

Summary

Family info

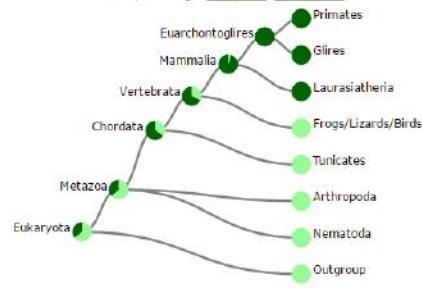
Name:	awaiting annotation
Accession:	TF337278
Description:	awaiting annotation
Taxonomic distribution:	Metazoa

Domain(s) and Function(s)

-  Enamelin (100% of seqs.)
-  HGNC: FVAM

Which species have awaiting annotation?

show percentage: by species/ by sequence



Legend: dark green shows present species/genes. Light green shows missing species/genes.

EMBL-EBI

Services

- By topic
- By name (A-Z)
- Help & Support

Research

- Overview
- Publications
- Research groups
- Postdocs & PhDs

Training

- Overview
- Train at EBI
- Train outside EBI
- Train online
- Contact organisers

Industry

- Overview
- Members Area
- Workshops
- SME Forum
- Contact Industry programme

About us

- Overview
- Leadership
- Funding
- Background
- Collaboration
- Jobs
- People & groups
- News

Data Warehouses Again

e!Ensembl BLAST/BLAT | BioMart | Tools | Downloads | Help & Documentation | Blog | Mirrors

Search: All species ▾ for Go

e.g. BRCA2 or rat 5:62797383-63627669 or rs699 or coronary heart disease

Browse a Genome
The Ensembl project produces genome databases for vertebrates and other eukaryotic species, and makes this information freely available online.

Popular genomes

 Human GRCh38.p6	 Human GRCh37
 Mouse GRCm38.p4	 Zebrafish GRCz10

★ Log in to customize this list

All genomes

View full list of all Ensembl species

Other species are available in [Ensembl Protocols](#) and [Ensembl Genomes](#)

New! From the 24th March – 5th May you can take part in our [interactive online webinar course](#), and afterwards catch up the recordings

Still using Human GRCh37? [Go to GRCh38](#)

Variant Effect Predictor 

Gene expression in different tissues 

Find SNPs and other variants for my gene 

Retrieve gene sequence 

Compare genes across species 

Use my own data in Ensembl 

ENCODE data in Ensembl 

What's New in Ensembl Release 84 (March 2016)

- 20 haematopoietic primary cell epigenomes from the BLUEPRINT project
- Mouse: update to Ensembl-Havana GENCODE gene set
- Track hub registry interface
- dbSNP 146 for Human, Cow and Dog
- Pairwise LD calculation on LD variant page

[Full details](#) | [All web updates by release](#) | [More news on our blog](#)

• 02 Jun 2016: [What's coming in Ensembl release 85](#)

• 25 Apr 2016: [DNA day and Malaria day: a story of scientific endeavour](#)

• 31 Mar 2016: [Ensembl 85 and Ensembl Genomes 32](#)

[Go to Ensembl blog](#)

Tweets by @ensembl

e! Ensembl @jonsenmbi Studying a dwarfism mutation in the PNKP gene #UsingEnsembl gene annotation buff.ly/24kBCE

e! Ensembl @jonsenmbi .dzerbino is at the @KeystoneSymp on human variation talking about Ensembl functional annotation now buff.ly/1O4ySMK

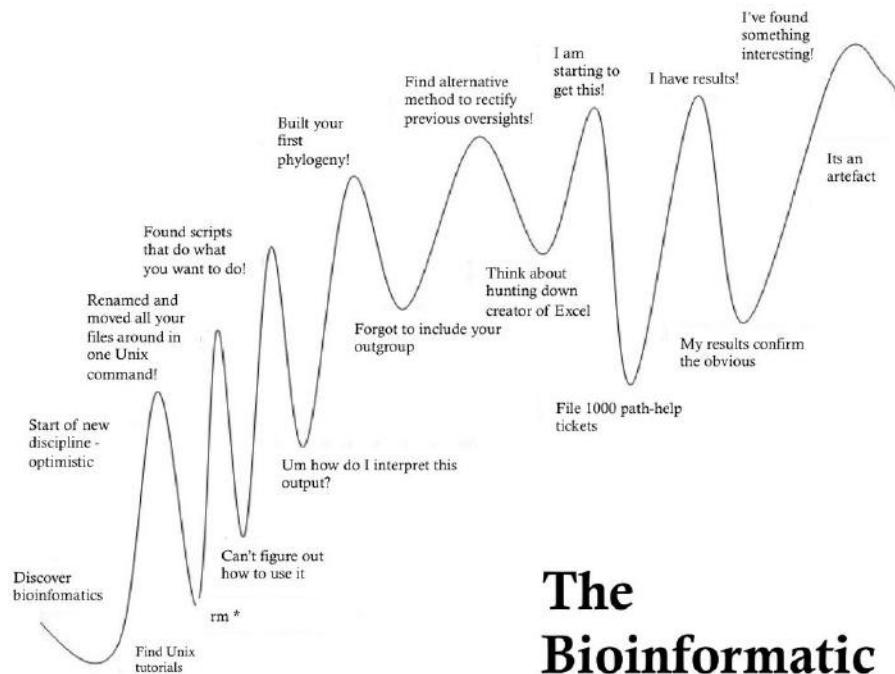
Embed View on Twitter

Ensembl is a joint project between EMBL-EBI and the Wellcome Trust Sanger Institute to develop a software system which produces and maintains automatic annotation on selected eukaryotic genomes.
www.ensembl.org/info/website/tutorials/orm37.html

Includes a list of additional current and previous funding bodies. How to cite Ensembl in your own publications.

EMBL-EBI

Learning Curve



**The
Bioinformatic
learning curve**

Q & A

Further reading

Swofford et al. 1996. Phylogenetic Inference. In Hillis, Moritz, & Mable [Eds.],

Molecular Systematics. Sinauer Associates, Sunderland, M.A.

Foster, 2007. Inferring phylogenetic relationships from sequence data. In Dear [Ed.],

Bioinformatics. Scion.

Lewis, 2001. Phylogenetic systematics turns of a new leaf. *Trends in Ecology and*

Evolution, 16: 30-37.

Felsenstein 2004. *Inferring Phylogenies*. Sinauer Associates, Sunderland, M.A.